1989

A Probabilistic Model For The Distribution Of Authorships And A Measure Of The Degree Of Research Collaboration

Isola S. Ajiferuke

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

A PROBABILISTIC MODEL FOR THE DISTRIBUTION OF AUTHORSHIPS AND A MEASURE OF THE DEGREE OF RESEARCH COLLABORATION

by

Isola S.Y. Ajiferuke

School of Library and Information Science

Submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario

November, 1988

ISBN 0-315-49326-7

ABSTRACT

The collaborative coefficient (CC), a measure that combines some of the merits of two earlier measures of research collaboration, is presented. This measure is used to compare the degrees of collaboration in the fields of engineering sciences, medical sciences, physical sciences, mathematical sciences, social sciences, and humanities. A theoretical model for the distribution of authorships is also developed. This model, the shifted Waring distribution, and 15 other discrete probability models are tested for goodness-of-fit against 96 data sets collected from the six fields listed above. The shifted inverse Gaussian-Poisson is found to be the best model. It is suggested that this model could be used in the estimation of the number of entries in an author index and in determining the maximum number of authors per paper to be included in an author index. A relationship is established between the parameters of this model and the collaborative coefficient.

1

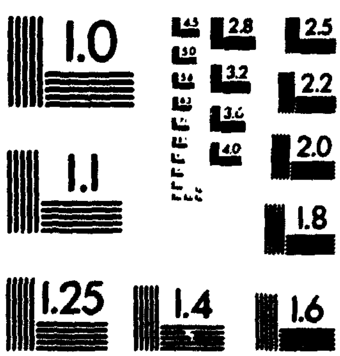


TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Bibliometricians have developed models for various social phenomena. The three most common models are Bradford's law (1934) for journal productivity, Lotka's law (1926) for scientific productivity, and Zipf's law (1949) for frequency of words in text. Models that have been established for other social phenomena are listed in articles by Simon (1955), Kendall (1961), Haitun (1982a), and Chen and Leimkuhler (1986).

One social phenomenon for which no model has been developed is collaboration in research, though a few have been suggested without their goodness-of-fit being tested. Just as the number of papers published in scholarly journals is taken as an indication of productivity, co-authorships of papers in a journal is often taken as an indication of collaboration in research.

Price and Beaver (1966), using memos circulated among members of an invisible college, inferred a Poisson model for the distribution of authorships, while Haitun (1982b), using the same data, classified the distribution of authorships among those stationary scientometric distributions that cannot be approximated by the Zipf distribution.

Worthen (1978) observed that for drug literatures the number of papers with one author makes up about 1/3 of the literature, those with two authors about 1/3 of the remaining literature, those with three, 1/3 of the remainder, and so on. Goffman and Warren (1980), on the other hand, observed that for the schistosomiasis literature, the number of publications with one author makes up about 2/3 of the literature, those with two authors about 2/3 of the remaining literature, those with three, 2/3 of the remainder, and so on. The two observations implied a geometric model for the distribution of authorships.

To compare the extent of collaboration in two or more fields (or periods of time), many studies have used either the mean number of authors per paper, termed the collaborative index (CI) by Lawani (1980), and/or the proportion of multiple-authored papers, called the degree of collaboration (DC) by Subramanyam (1983a), as a measure of the strength of collaboration in a discipline.

While CI differentiates among levels of authorships and is very easy to calculate, it is not easily interpretable as a degree for it has no upper limit (i.e., it is not expressible in terms of percentage) and it gives a non-zero weight to single-authored papers, which involve no collaboration. DC, on the other hand, is easy to calculate, easily interpretable as a degree, and gives zero weight to single-authored papers but it does not differentiate among levels of authorships.

The objectives of this research, therefore, are to :

(i) Derive a theoretical model for the distribution of authorships. (ii) Test the derived model and some other discrete probability models (for goodness-of-fit) against empirical data.

(iii) Derive a measure of the degree of collaboration in research that combines some of the merits of both CI and DC.

(iv) Use the derived measure to compare the degrees of collaboration in some fields.

(v) Determine whether or not the degree of collabo.ation has been increasing over time in general or in specific fields.

(vi) Relate the measure of the degree of research collaboration to the parameters of the derived model.

CHAPTER 2

LITERATURE REVIEW

Since this study deals with research collaboration, a general review of the phenomenon is presented below. The review includes the merits and kinds of research collaboration as well as its relationship with other aspects of research.

2.1 Advantages and Disadvantages of Collaboration

Collaboration in scientific research dates back at least as early as the seventeenth century. The earliest collaborative paper, found by Beaver and Rosen (1978a), was published in 1665 and attributed to Hooke, Oldenburg, Cassini, and Boyle.

Collaboration in research can involve many forms of activity ranging from offering general advice and opinion to active and sustained participation and contribution of physical and intellectual resources. Hefner (1981) has characterized collaboration as being 'theoretical' (i.e., rendering advice, ideas, or criticism), or 'technical' (i.e., providing tangible assistance in a research endeavour). He also distinguished between 'co-authors' (i.e., those who share authorship of a publication), and 'sub-authors' (i.e.those who are not co-authors, but whose theoretical or technical assistance in the research project is acknowledged by the author(s) in a publication). However, the practice of naming co-authors varies widely and depends on the attitudes and perceptions of the individuals involved, and the policies and traditions of the organization in which the research endeavour takes place (Subramanyam, 1983a).

The degree of collaboration varies from one discipline to another. However, the multiplicity of authorship in different disciplines tends to remain relatively stable from country to country (Meadows, 1974). It is generally high in the intensively collaborative scientific fields, but low in the humanities, in which the lonely scholar, working without the trappings of 'big science', still produces much of the literature (Subramanyam, 1983a). Because of the time-consuming nature of data collection, data processing and analysis, there tends to be more collaboration in 'experimental' research than in 'theoretical' research.

The reasons and motives that have been adduced for research collaboration include :

(i) Professionalization (i.e., a process which organizes a group of individuals along a set of attributes) of science (Beaver & Rosen, 1978a).

(ii) Attempts to reduce marginal per author cost of publishing (Schweser, 1983). The monetary costs of academic publishing are to a large extent absorbed by the writer's university. The costs incurred by the writer are mainly personal. These costs can be measured in terms of the time and effort expended for researching and writing the article, lost professional opportunities such as outside consulting fees, and foregone leisure time.

(iii) Opportunities to join resources and divide labour, alleviate academic isolation, sustain motivation through external commitments, and create interpersonal energy for the completion of projects (Fox & Faver, 1984).

(iv) Defence against the growing complexity of subject matter, provision of intellectual stimulation when different individuals focus on a single problem, and facilitation of synthesis of theoretical assumptions of different scientific disciplines (Patel, 1973).

In spite of all the advantages listed above, collaboration has its costs and disadvantages. Some of these are :

(i) Process costs : A primary cost of collaboration is the time it requires for negotiation and exchange (Fox & Faver, 1984). Any collaborative venture requires a certain amount of planning and management as well as exchange and assistance on its theoretical or technical aspects. In addition to the time expense, collaborators (especially long-distance collaborators) frequently incur expenses in telephone, mail, copying, and travel. The extent of these costs depends not only on the nature of the project but also on the extent and source of its funding. A different, but no less important, process cost is the personal, socio-emotional cost of developing and maintaining a good working relationship with one's collaborators. Any interpersonal relationship requires the investment of both time and emotional energy. And even in the best collaborative associations, there are inevitable differences of opinion.

(ii) Outcome costs : Collaboration can also be costly for the outcome of the project (Fox & Faver, 1984). First, a slow or recalcitrant partner can seriously delay or jeopardize a project. Delay is particularly hazardous for work in competitive research areas. In these cases, a sluggish collaborator can result in the project's findings being "scooped" or "anticipated" during the delay. Second, co-authorship may present a problem in the evaluation and allocation of credit for research and writing. This problem manifests itself in bibliometric studies of research productivity and citation analysis (Lindsey, 1980). First of all, it is not always easy to determine the magnitude of importance of each author's contribution to a paper from the ordering of the names of co-authors because while it is generally true that the name of the principal investigator is almost always mentioned first, the order in which the remaining co-authors are named in the paper does not necessarily reflect the degree of collaboration (Zuckerman, 1968). Secondly, do we award a credit or citation (for either publication or receiving a citation) to only the senior author of an n-authored paper, give one credit or citation to each co-author or award 1/n credits or citations to each co-author ? This problem is further magnified in cases where the authors are obliged by a particular journal's editorial policy to list themselves in alphabetical order by surname. Such a policy has been shown, by Over and Smallman (1973), to have resulted in limited publication of collaborative research in the Journal of Physiology, London, by scientists with surnames commencing p-z .

(iii) Extra cost of production of indexing and abstracting products: Bibliographers must list entries under several author names, and abstracting services must add extra index entries (McCauley, 1963). These would result in bulkier products, and hence, extra cost of production. And in cases like that of the Science Citation Index, where indexing service cannot afford to index more than the first author, all other collaborators are inaccessible.

2.2 Kinds of Collaboration

According to Subramanyam (1983a), depending on the participants, the following kinds of collaboration can be identified :

(i) Teacher-pupil collaboration.

Here, the professor in a university department provides the ideas and guidance, and sometimes also the funds from a research grant, and the research assistant or student does most of the bench work. Teacher-pupil collaboration is greater in the hard sciences than in social sciences and humanities. In a survey of 80 university graduate departments by Lodahl and Gordon (1972), chemists and physicists were found to collaborate with larger numbers of graduate students in research than sociologists and political scientists.

(ii) Collaboration among colleagues.

Here, we refer to the kind of collaboration that occurs between colleagues in a university (either within a department, a faculty or the university as a whole), corporate research centre, or invisible college. The collaboration can be within discipline or between disciplines. In interdisciplinary fields such as environment, energy, or space research, scientists and engineers from a wide variety of specialties often collaborate. It is not uncommon for chemists, chemical engineers, mathematical engineers, biophysicists, and other specialists to be working together in an interdisciplinary project. Husband-and-wife teams can also be included in this category.

(iii) Supervisor-assistant collaboration.

This is the kind of collaboration that occurs between a principal investigator in research projects requiring extensive use of laboratory facilities or very specialized equipment and an array of laboratory assistants and technicians. In the past, such research assistants might have their assistance acknowledged in the body of a paper but nowadays, some of them are regarded as junior colleagues (Meadows, 1974).

(iv) Research-consultant collaboration.

This is the kind of collaboration that occurs between an individual researcher or a research team and a consultant or a consulting firm. In most cases, the assistance of the consultant or consulting firm is sought for data collection, data processing and analysis.

(v) Inter-organizational collaboration.

Inter-organizational collaboration is often necessitated by mutual interests or by the complexity of a research project, or when researchers in one organization may need to use expensive equipment or specialized service available at another organization.

(vi) International Collaboration.

International collaboration in research can take many forms : the sharing of unique data sources, correspondence by mail, exchanges of ideas at conferences, visits to foreign laboratories or universities (especially scholars on sabbatical leave), and the exchange of papers are all common examples of collaborative activity. Frame and Carpenter

(1979), in a study of international co-authorship patterns in nine scientific fields, found that :

(1) The more basic the field, the greater the proportion of international co-authorships, i.e., the degree of international collaboration in such basic fields as physics, mathematics, and chemistry is higher than in applied fields such as engineering and technology, clinical medicine and biomedical research.

(2) The larger the national scientific enterprise, the smaller the proportion of international co-authorship.

(3) Extra-scientific factors, such as geography, politics, and language, play a strong role in determining who collaborates with whom in the international scientific community.

It should be noted here that it is possible for the research-consultant and inter-organizational kinds of collaboration to occur at the international level.

2.3 Bibliometric Studies of Research Collaboration

Collection of data in studies of research collaboration is a difficult problem. Because of the complex nature of interactions that take place between or among collaborators over a period of time, the precise nature and magnitude of collaboration cannot be easily determined by the usual methods of observation, interviews or questionnaires (Subramanyam, 1983a).

If we assume that (i) all collaborative research effort results in one or more published papers, (ii) all the collaborators are mentioned as co-authors in the publications, and (iii) all the co-authors mentioned in the publications actually collaborated in the research effort, then bibliographic data containing the names of all the co-authors offer an unobtrusive indicator for the study of collaboration (Subramanyam, 1983a).

It should be noted here that these assumptions are not always true. For example, Edge and Mulkay (1973), in their study of the development of radio astronomy in Britain, found several instances where researchers had collaborated closely, but had written up the results as separately-authored papers (this was particularly true of collaborations between experimentalists and theorists).

However, despite the limitations, the bibliometric method is still considered to be superior to traditional methods of survey, observation, interviews or questionnaires. According to Subramanyam, the principal advantages of using the number of co-authors as a measure of collaboration are that it is

(i) invariant ,

(ii) easily and inexpensively ascertainable ,

(iii) quantifiable , and

(iv) non-reactive (i.e., the process of ascertaining collaboration does not affect the process of collaboration itself).

The bibliometric method facilitates the investigation of the relationship between research collaboration and variables pertaining to the research problem and the research environment, by applying statistical techniques such as regression, correlation, and factor analysis. A review of some of the bibliometric studies of research collaboration that have been reported in the literature is presented below.

2.3.1 Relationship between Collaboration and Financial Support

A few studies have shown that external financial support for research is associated more often with collaborative research effort than with solo research.

In their study of articles published in the **American Sociological Review** and **Social Forces**, Hirsch and Singleton (1964) found that supported articles in both journals had more authors per article than non-supported articles. This finding was corroborated by data for four sociological journals used in Patel's study of collaboration in the growth of sociology (1973).

Hefner (1981) also studied the relationship between collaboration and financial support for research in four disciplines (political science, psychology, biological science, and chemistry). In all of these disciplines, financial support for research was associated with an increase in the total number of persons (including co-authors and sub-authors) involved in the production of knowledge per research paper ; the association was particularly strong in biological science and chemistry.

In 1982 Subramanyam and Stephens also investigated the relationship between collaboration and funding, using research papers in biochemistry and chemical engineering. They found a higher proportion of multiple-authored papers than single-authored papers was supported by grants.

Funding per se does not necessarily 'cause ' an increase in collaborative research, for it is possible that the more people involved in a research project, the more they might have need for additional funding (Hefner, 1981). However, with financial support, the researcher may have greater access to complicated, expensive equipment, which in turn, may require a large number of people to operate and utilize. Furthermore, the researcher may now be able to pay others to do the tasks that he himself might otherwise have done.

2.3.2 Relationship between Collaboration and Productivity

Studies by Price and Beaver (1966), Zuckerman (1967), Beaver and Rosen (1979), Pao (1982), and Pravdic and Oluic-Vukovic (1986) have shown a strong relationship between collaboration and productivity.

In their study of collaboration in Information Exchange Group No. 1 (on oxidative phosphorylation and terminal electron transport, organized by the National Institutes of Health), Price and Beaver found that productivity increased as the number of collaborators increased. They noted the existence of a small core of extremely active researchers, surrounded by a large floating population of people who collaborated with the leaders on only one or two projects and then disappeared. Price and Beaver suggested that "part of the social function of collaboration is that it is a method of squeezing papers out of the rather large population of people who have less than a whole paper in them ."

Zuckerman's study of 41 Nobel laureates also showed a high degree of correlation between collaboration and productivity. In general, laureates published more and were more apt to collaborate than a matched sample of scientists.

In Beaver and Rosen's study of the nineteenth century French scientists, scientists who collaborated were found to exhibit higher productivity than those who did not. Beaver and Rosen noted further that a scientist whose first paper was jointly authored stood a better chance of producing another paper than did one who wrote his first paper alone. If the first paper was collaborative, the probability of producing another paper was 0.8 while writing one's first paper alone reduced the probability to 0.5.

Pao investigated the relationship between collaboration and productivity in musicology, a 'humanistic subject'. Though only 15% of the literature of musicology was the result of collaborative authorship, a few heavy collaborators were found to be highly prolific.

Pravdic and Oluic-Vukovic, in their study of the most prolific Croatian chemists, found that not all the effects of scientific collaboration could be explained by a simple linear correlation between the number of papers produced by scientists and the number of their collaborators (the coefficient of determination was 0.4435). They, however, noted that the higher the value of the "collaboration measure" (this measure, defined as the total number of co-authors appearing in the joint papers divided by the total number of individual collaborators, refers to the frequency of collaboration among the same authors), the greater is the impact on the author's productivity.

2.3.3 Relationship between Collaboration and Quality of Research

Gordon (1980) found a significant relationship between the levels of multiple authorship of papers to an astronomy journal, and their frequency of acceptance for publication ; the larger the number of authors for a paper, the higher its chance of acceptance for publication. One possible explanation for this observation, according to Gordon, is that multiple authorships provide the opportunity for cross-checking and pre-submission internal refereeing. He further suggested that the larger the number of authors for a paper, the greater the likelihood that it is predominantly observational, rather than theoretical. And the more observational a paper is, the more clearly delineated and "shared" are the criteria which are considered to be appropriate for its evaluation.

Presser (1980) also analyzed whether editorial judgement (based on referee evaluations) on papers submitted to a leading social psychology journal is related to the number of authors for a paper. He found the percentage of multiple-authored papers accepted for publication to be higher than that of single-authored papers. Presser's own explanation for this observation is that collaboration may increase efficiency through the principle of the division of labour. Thus two investigators - say, one particularly skillful in experimental design, the other in data analysis - should produce a better paper than either working alone.

Lawani (1986) used a different method to assess the quality of papers in his own investigation of the relationship between collaboration and quality. He collected three samples of cancer research literature representing three levels of quality. Journal articles abstracted in the

Year Book of Cancer constituted his first-order quality sample (the assumption was that these articles are normally selected by distinguished editors who critically review more than 500,000 journal articles each year), articles listed but not abstracted in the Year Book of Cancer for reasons of lack of space constituted the second-order quality sample while cancer research papers published in Biology Abstracts but not abstracted or listed in the Year Book of Cancer constituted the average-order or random sample.

Lawani sorted the papers according to the number of authors and found that as the number of authors per paper increases, the proportion of high quality papers also increases. He also compared the average number of authors per paper for the three different samples, and found the average number of authors per paper for the random sample to be significantly lower than that of first-order quality sample and also lower than that of the second-order quality sample.

2.3.4 Relationship between Number of Authors and Title Length of Journal Articles

Kuch (1978), using several issues of the American Journal of Physiology, Biochimica et Biophysica Acta, Chest, Journal of the American Society for Information Science, and Biochemistry, found a positive correlation between the number of authors of papers in the first four journals and the number of significant words in the titles of the same papers. The journal **Biochemistry**, however, showed no correlation, the data in this case being compatible with the null hypothesis that no relation exists between title length and number of authors.

Two possible explanations of the positive relation between title length and number of authors, given by Kuch, are :

(1) multiple-authored papers are based on more extensive studies than single-authored papers, hence communicate more information , hence tend to have longer titles, and

(2) interaction among co-authors acts to increase title length : each author may have a different idea of why the paper is significant, and there seem to be no forces at play which act to limit the length of titles. The title grows until each co-author is satisfied that the particular term or phrase that suits him is included. The greater the number of co-authors, the more terms or phrases may be included.

2.3.5 Relationship between Collaboration and Impact

Studies investigating the relationship between collaboration and impact have so far shown little or no differences in the impact of single-authored versus multiple-authored papers in various disciplines.

Oromaner (1975) investigated this relationship using the number of citations received by a paper as a measure of impact. He selected all single-authored articles published in three prominent sociological journals - **American Journal of Sociology, American Sociological Review,** and **Social Forces** - and a sample of all multiple-authored articles appearing in the journals during 1959-1961. He then counted the citations made to these articles in all full-length articles published in the three journals for a period of ten years. Self-citations were eliminated and each article was given one credit for each article in which it was cited regardless of the number of times it was cited in that article. It was found that :

(1) single-authored articles are somewhat less likely to have an impact than are multiple-authored articles, and

(2) of those articles which do have an impact, the single-authored articles are somewhat more likely to have low impact than are the multiple-authored articles.

Lindsey (1978) investigated the same relationship for research papers in six academic disciplines (biochemistry, economics, psychiatry, psychology, social work, and sociology) but found no overall difference in the number of citations received by single-authored and multiple-authored papers. Lindsey did report, however, that there appeared "to be a consistent proportional increase in citations attracted and number of authors" in the two applied fields, social work and psychiatry. Lindsey, therefore, suggested that collaboration and impact might be positively correlated in the applied fields.

Bayer (1982), in his assessment of the applied area of family science, found no relationship between collaboration and citation frequency.

In order to verify the true relationship between collaboration and impact for applied fields, Smart and Bayer (1986) analyzed a ten-year citation rates of 270 randomly selected articles in three applied fields - clinical psychology, educational measurement, and management science. The mean number of citations per paper (self-citations included or excluded) was found to be lower for single-authored than for multiple-authored papers in these three applied fields. However, these differences are not statistically significant for articles in clinical psychology or in educational management ; only multiple-authored articles in management show a statistically significant higher rate.

An entirely different kind of study was carried out by Diamond, who used longitudinal data for the Berkeley Mathematics Department on salaries and citations by authorship category to investigate the weight employers attach to various categories. The effect of citations on earnings was estimated using ordinary-least-squares regression analysis with the natural logarithm of salary as the dependent variable and various independent variables that have traditionally been thought to influence earnings. The analysis showed that a citation to a single-authored article is worth less to its author than a citation to a multiple-authored article. Diamond, therefore, suggested that if the finding turns out to be true for other disciplines (as well as for different samples of mathematicians), one explanation worth considering would be that citations to multiple-authored articles are a proxy for the trait of collegiality that is rewarded by departments in the determination of salaries.

2.3.6 Multiple-Authorship Trends

Multiple authorship has been increasing in practically all fields in recent decades (Price, 1963 ; Patel, 1972 ; Balog, 1980 ; Gladding, 1980 ; Utz, 1981 ; Mendenhall & Higbee, 1982 ; Over, 1982). The only field that has not been investigated is the humanities. This might be due to the general notion that the field is non-collaborative in nature.

Although the extent of collaboration in research has steadily increased, individual research effort is nowhere near extinct (Subramanyam, 1983a). Price (1963), using data from **Chemical Abstracts**, 1910-1960, noted that the proportion of multiple-authored papers had accelerated steadily since the beginning of the twentieth century, and that if the same trend continued, there would be no single-authored paper by 1980. Obviously, this has not happened and Price's contention has been challenged by a few scholars (Clarke, 1964 ; Hodder, 1980). According to Hodder, it seems that after initial recognition of the value of collaborative research, scientists tend to assert their individuality again.

As noted earlier, in section 2.1, the extent of collaboration varies from discipline to discipline. The extent of variation between two or more disciplines has been investigated in many studies (Mitchell, 1961 ; Stefianiak, 1982 ; Subramanyam & Stephens, 1982 ; Pao, 1982 ; Subramanyam, 1983b), using objective measures of the degree of collaboration. The two commonly used measures are the mean number of authors per paper, and the proportion of multiple authored papers.

CHAPTER 3

DATA COLLECTION

Most bibliometric studies of research collaboration have used either scholarly journals or abstracting journals or both as their sources of data. Because of their comprehensive nature, abstracting journals were used as the source of data for this study. However, only research papers published in scholarly journals or presented at conferences were counted. Books, dissertations, theses, and research papers by anonymous or corporate authors were excluded because :

(1) Some books are made up of several articles, which are at times written by different authors, and not all the abstracting journals cover monographic literature. However, book chapters or parts of books abstracted separately were counted.

(2) Dissertations and theses are of single authorships by nature.

(3) One cannot easily ascertain the number of people responsible for research works credited to anonymous or corporate authors. Also, editors, compilers, translators, and moderators of symposia were not regarded as co-authors since they are not intellectually responsible for the research works attributed to them.

In order to see how well the probability models fit the distribution of authorships in various fields, data was collected from engineering sciences, medical sciences, physical sciences, mathematical sciences, social sciences, and humanities.

Engineering Index abstracts literature from all the disciplines in engineering sciences while Index Medicus does the same for the disciplines in medical sciences. However, no single abstracting journal covers all the disciplines in either the physical sciences, mathematical sciences, social sciences, or humanities. Hence, a few abstracting journals were chosen to represent each of these fields (papers from the abstracting journals representing each field were not added together because of journal overlap). Availability, scope, and prominence are the factors taken into consideration in choosing the abstracting journals. The chosen abstracting journals are listed below in Table 1.

In order to compare the extent of collaboration in recent decades, data was collected at 5-year intervals. Because some of the chosen abstracting journals came into being late in the fifties or early in the sixties, the base year is chosen to be 1960. The years chosen by systematic sampling are 1961, 1966, 1971, 1976, 1981, and 1986. However, data for philosophy will be collected for 1967/68, 1969, 1971, 1976, 1981 and 1986 because the first issue of Philosopher's Index was published in 1967/68.

The manual versions of these journals were used because most of them list all the authors of each paper whereas a maximum of 4 authors per paper are usually given by most of the online versions. In a few cases where the list of authors was truncated, the original source of the paper was consulted.

Systematic sampling method was used to collect data from **Engineering Index** and **Index Medicus** while cluster sampling method was employed in collecting data from **Chemical Abstracts**, **Biological Abstracts** and **Physics Abstracts**, i.e., a few monthly or weekly issues were selected at random for each year and all the papers in each issue were counted. For the others, which index a relatively small number of records per year, all the papers for each year were counted. In this case, the papers in an abstracting journal for a particular year are being used as a sample of all the papers published that year in the corresponding discipline. While the selection of papers for inclusion in an abstracting journal might not be random, there is no reason to believe that the distribution of authorships for those included is different from the distribution for those excluded.

The papers selected for each year for each of the abstracting journals were sorted into different levels of authorships. The number of papers at each level were then counted and the resulting distribution of authorships taken as one data set. Hence, a total of 96 data sets were used in the analysis (See Appendix A for the data sets).

Table 3.1 : List of the Abstracting Journals.

Field	Abstracting Journal
1. Engineering Sciences	Engineering Index (1917-
2. Medical Sciences	Index Medicus (1960-
3. Physical Sciences	
	Physics Abstracts (1903-
	Chemical Abstracts (1924-
	Biological Abstracts (1926/7
4. Mathematical Sciences	
	Mathematical Reviews (1940-
	Statistical Theory & Methodology Abstracts (1960-
	Computer Abstracts (1957-
5. Social Sciences	
	Sociological Abstracts (1953-
	International Political Science Abstracts (1951-
	Key to Economic Science and Managerial Sciences (1953-
	Psychological Abstracts (1922-
	Library and Information Science Abstracts (1950-
6. Humanities	
	Historical Abstracts (1955-
	Abstracts of English Studies (1958-
	Philosopher's Index (1967/68-

CHAPTER 4

MEASURES OF COLLABORATION

An objective measure of the strength of collaboration in a discipline is very useful in comparing the extent of collaboration in disciplines or periods of time. In this chapter, existing measures are reviewed and a new one is derived. The new measure is used to compare the extent of collaboration in six fields and also to determine whether collaboration has been increasing over time in general or in specific fields.

In the following, we write

j = the number of authors for a research paper ;

f_j = the number of j-authored research papers published in a discipline during a certain period of time ;

N = the total number of research papers published in a discipline during a certain period of time ; and

k = the greatest number of authors per paper in a discipline during a certain period of time, i.e., j_{max}.

4.1 Present Measures

(i) Collaborative Index (CI)

$$CI = \frac{1}{N} \sum_{j} j \, f_j$$

$$= \text{mean number of authors per paper.}$$

CI differentiates among levels of authorships and is very easy to calculate but it has the following disadvantages :

(1) it is not easily interpretable as a degree for it has no upper limit (i.e., it is not expressible in terms of percentage).

(2) it gives a non-zero weight to single-authored papers, which involve no collaboration.

(3) it is easily affected by extreme values.

One way of avoiding these problems is to use 1 - 1/CI as a measure of collaboration. However, this measure has no theoretical base, though very easy to calculate.

(ii) Degree of Collaboration (DC)

$$DC = 1 - \frac{f_1}{N}$$

DC is easy to calculate and easily interpretable as a degree (for it lies between zero and one), not easily affected by extreme values, gives zero weight to single-authored papers, and always ranks higher a discipline (or period of time) with a higher percentage of multiple-authored papers. However, DC does not differentiate among levels of multiple authorships. For example, in Table 4.1, 1961 and 1966 have almost the same proportion of multiple-authored papers but different distributions of multiple-authored papers. However, this difference in distributions is not reflected in their DC values (see Table 4.2).

(iii) CI and DC

According to Subramanyam (1983a), presenting both CI and DC gives a fairly clear idea of the extent of collaboration in a discipline. This is true in some cases, although problems may arise in comparative studies.

There are five kinds of situation that can occur when using both CI and DC to compare the degrees of collaboration in two disciplines (or two periods) r & s :

(1) $CI_r = CI_s$ and $DC_r = DC_s$;

(2) $CI_r = CI_s$ but $DC_r \neq DC_s$;

(3) $CI_r \neq CI_s$ but $DC_r = DC_s$;

(4) $CI_r > CI_s$ and $DC_r > DC_s$ or vice-versa (e.g., for 1986 and 1976 in Table 4.2) ; and

(5) $CI_r > CI_s$ but $DC_r < DC_s$ or vice versa (e.g., for 1966 and 1961 in Table 4.2).

In the first and fourth situations, the conclusions are obvious. In the other situations, however, there are

Table 4.1 : Distribution of Authorships for Library and Information Science

Number of Authors	Number of Papers (% in Parentheses) 1961	1966	1971	1976	1981	1986
1	783 (94.11)	1021 (94.28)	1968 (86.35)	2771 (87.06)	3697 (83.47)	4971 (82.88)
2	43 (5.17)	48 (4.43)	232 (10.18)	312 (9.80)	559 (12.62)	786 (13.10)
3	6 (0.72)	10 (0.92)	54 (2.37)	65 (2.04)	123 (2.78)	170 (2.83)
4	-	3 (0.28)	15 (0.66)	23 (0.72)	33 (0.75)	36 (0.60)
5	-	1 (0.09)	8 (0.35)	6 (0.19)	8 (0.18)	17 (0.28)
6	-	-	1 (0.04)	5 (0.16)	5 (0.11)	10 (0.17)
7	-	-	0 (0.00)	1 (0.03)	4 (0.09)	3 (0.05)
8	-	-	1 (0.04)	-	-	0 (0.00)
9	-	-	-	-	-	2 (0.03)
10	-	-	-	-	-	3 (0.05)
Total	832	1083	2279	3183	4429	5998

Table 4.2 : Measures of Collaboration for Library and Information Science

Year	CI	DC	CC
1961	1.0660	0.0590	0.0306
1966	1.0748	0.0570	0.0311
1971	1.1880	0.1365	0.0752
1976	1.1778	0.1294	0.0711
1981	1.2224	0.1653	0.0904
1986	1.2356	0.1712	0.0938

conflicts between the two measures and it is not clear on which one to base our judgement. In order to avoid this problem, we prefer to use a single measure that incorporates some of the merits of both of the above.

4.2 Proposed Measure

(i) Derivation

Imagine that each paper carries with it a single "credit", this credit being shared among the authors. Thus if a paper has a single author, the author receives one credit; if 2 authors, each receives 1/2 credits and, in general, if we have x authors, each receives 1/x credits. This is the same as the idea of fractional productivity defined by Price and Beaver (1966) as the score of an author when he is assigned 1/n of a unit for one item for which n authors have been credited. Hence, the average credit awarded to each author of a random paper is E[1/X], a value which lies between 0 and 1. If 0 is to correspond to single authorship, then we define the collaborative coefficient, CC, as

$$\begin{aligned} CC &= 1 - E[1/X] \\ &= 1 - \sum_j (1/j) \; P(X=j) \end{aligned}$$

and its sample estimate is

$$1 - \frac{f_1 + (1/2)\, f_2 + \ldots + (1/k)\, f_k}{N}$$

$$= \quad 1 \; - \; \frac{\Sigma\,(1/j)\,f_j}{N}$$

where f_j, N and k are as previously defined.

Note that :

(i) $0 \le CC \le 1$;

(ii) $CC \to 0$ as $P(X=1) \to 1$ and $CC \to 1 - 1/j$ as $P(X=j) \to 1$, i.e., CC tends to zero as single-authored papers dominate and to $1 - 1/j$ as j-authored papers dominate. The behaviour of CC as the maximum number of collaborators $k \to \infty$ depends on the form of f_i ;

(iii) $CC = 1 - E[1/X]$

$$= 1 - \Sigma\, 1/j\; P(X=j) \le 1 - P(X=1) = DC \le 1 \le CI,$$

i.e, $CC \le DC \le CI$; and

(iv) by Jensen's inequality, $E[1/X] \ge 1/E[X]$ so that $1 - E[1/X] \le 1 - 1/E[X] \Rightarrow CC \le 1 - 1/CI$

It is worth remarking that a collaborative coefficient can be calculated for a particular author, in which case the papers to which he or she contributed would constitute the sample of papers. Hence, the degree of collaboration of an individual can be compared with the standard in his or her discipline.

(ii) Example

CC for the distribution of authorships for 1966 in Table 4.1 is calculated thus :

$$1 - \frac{(1 \times 1021)+(1/2 \times 48)+(1/3 \times 10)+(1/4 \times 3)+(1/5 \times 1)}{1083}$$

$$= 1 - \frac{1021 + 24 + 3.333 + 0.75 + 0.2}{1083}$$

$$= 1 - 1049.283/1083$$

$$= 1 - 0.9689$$

$$= 0.0311$$

Similarly, values of CC for 1961, 1971, 1976, 1981, and 1986 are calculated and displayed along with the corresponding values of CI and DC in Table 4.2.

On the basis of the CC values, collaboration among researchers in library and information science in 1966 is higher than in 1961, which is contrary to what is indicated by the DC values but in agreement with the ordering based on the CI values. The effect of the higher levels of multiple authorships in 1966 is stronger than the effect of the higher proportion of multiple authorships in 1961. For other pairs of years, the same ordering would be made on the basis of the three measures. However, the percentage difference may differ. For example, the percentage increase in the CC value from 1981 to 1986 is 3.7%, which is greater than 1.8% for the DC value because the DC value did not reflect the higher levels of multiple authorships in 1986.

4.3 Collaboration Trends

The values of the various measures of collaboration for all the disciplines are given in Appendix B. These values are plotted by field and year in Figures 4.1 to 4.18. It is observed that :

(i) Engineering Sciences

The average number of authors per paper increased from 1.5 in 1961 to 2.3 in 1986 while the percentage of multiple-authored papers increased from 36% in 1961 to 64% in 1986. The CC value of 0.2 in 1961 is found to have doubled by 1986.

(ii) Medical Sciences

The average number of authors per paper increased from 1.8 in 1961 to 3.2 in 1986. The percentage of multiple-authored papers increased from 48% in 1961 to 79% in 1986 and by that year the most frequent number of authors per paper has changed from one to two (see Table A.2 in Appendix A). The CC value is found to have increased from 0.28 in 1961 to 0.54 in 1986.

(iii) Physical Sciences

The average number of authors per paper of about 1.83 in 1961 is found to have increased to about 2.96 in 1986. The percentage of multiple-authored papers of less than 61% increased to more than 75% in 1986. It is noted that by 1976 the most frequent number of authors per paper has changed from one to two (see Tables A.3 to A.5 in Appendix A). The CC value also increased from about 0.3 in 1966 to about 0.5 in 1986.

(iv) Mathematical Sciences

The average number of authors per paper increased slightly from about 1.26 in 1961 to about 1.59 in 1986 while the percentage of multiple-authored papers increased from less than 74% in 1961 to more than 34% in 1986. The CC value increased from about 0.1 in 1961 to about 0.24 in 1986.

(v) Social Sciences

The average number of authors per paper increased slightly from about 1.21 in 1961 to about 1.46 in 1986 while the percentage of multiple-authored papers increased from about 16% in 1961 to about 30% in 1986. The CC value increased from about 0.09 in 1961 to about 0.17 in 1986.

(vi) Humanities

This field is still dominated by single-authored papers. The average number of authors per paper is found to be less than 1.1 in 1986 with at least 93% of the papers being single-authored. The CC value is also found to be just about 0.024 in 1986.

FIG. 4.1 : CI VERSUS YEAR
ENGINEERING SCIENCES

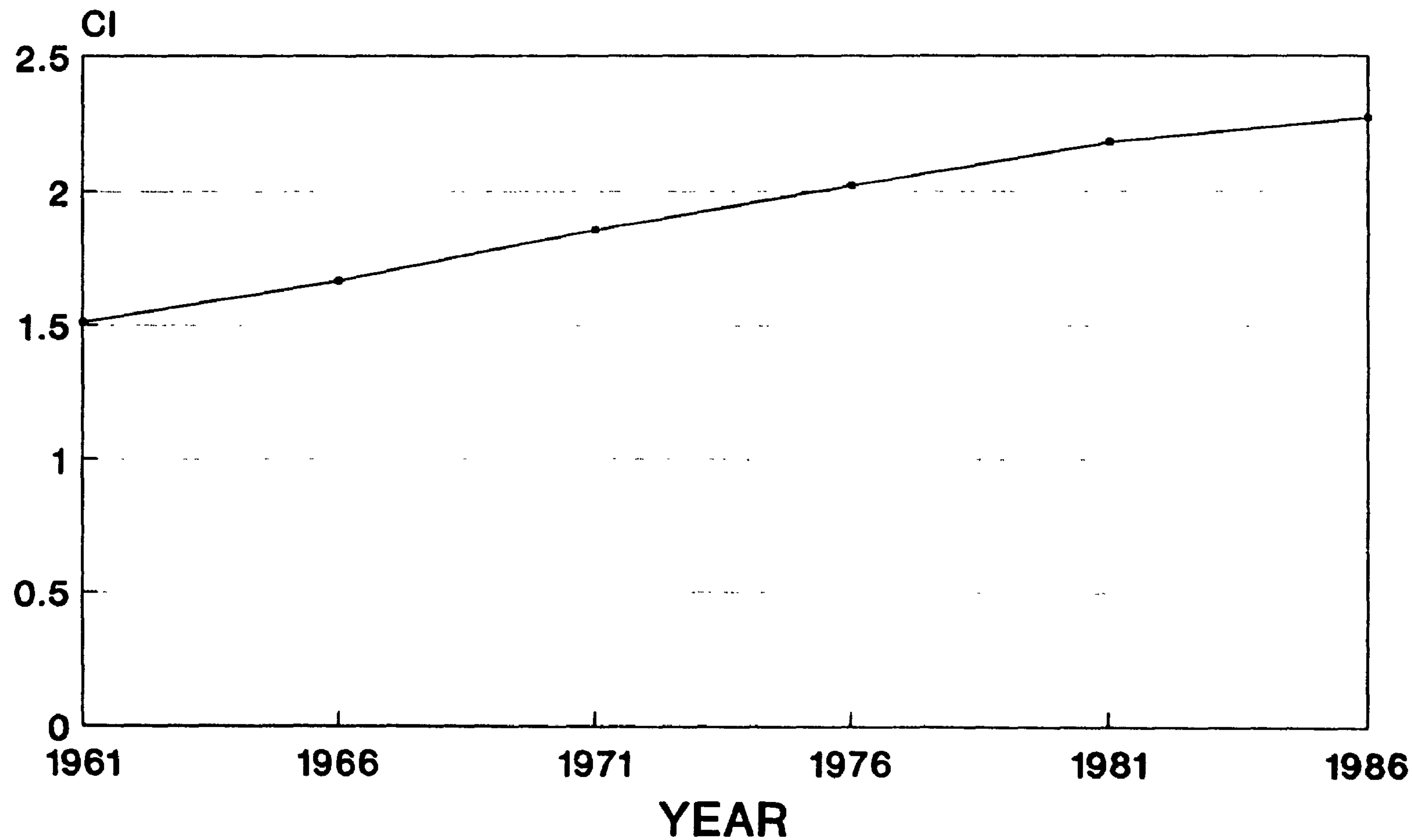


FIG. 4.2 : CI VERSUS YEAR
MEDICAL SCIENCES

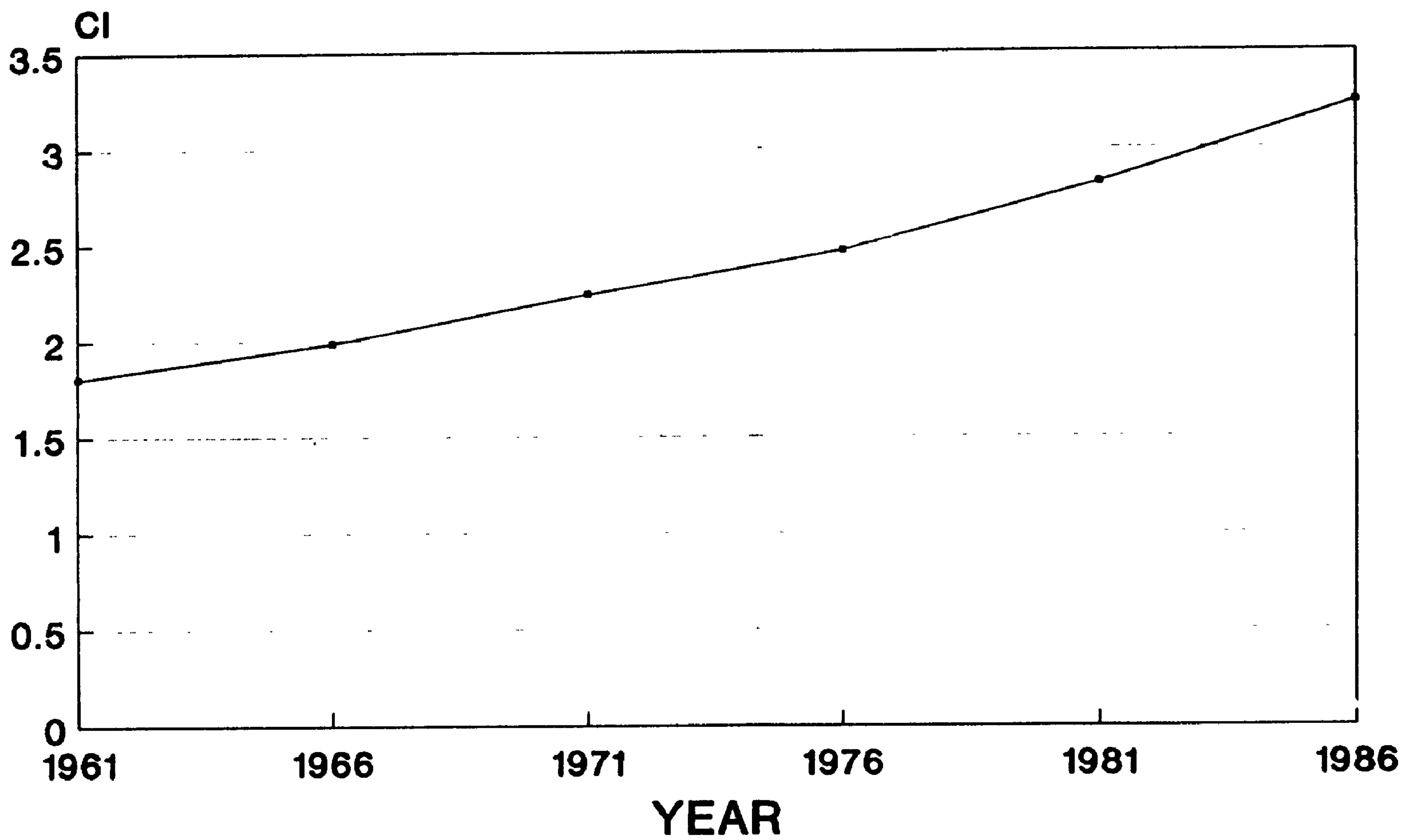


FIG. 4.3 : CI VERSUS YEAR
PHYSICAL SCIENCES

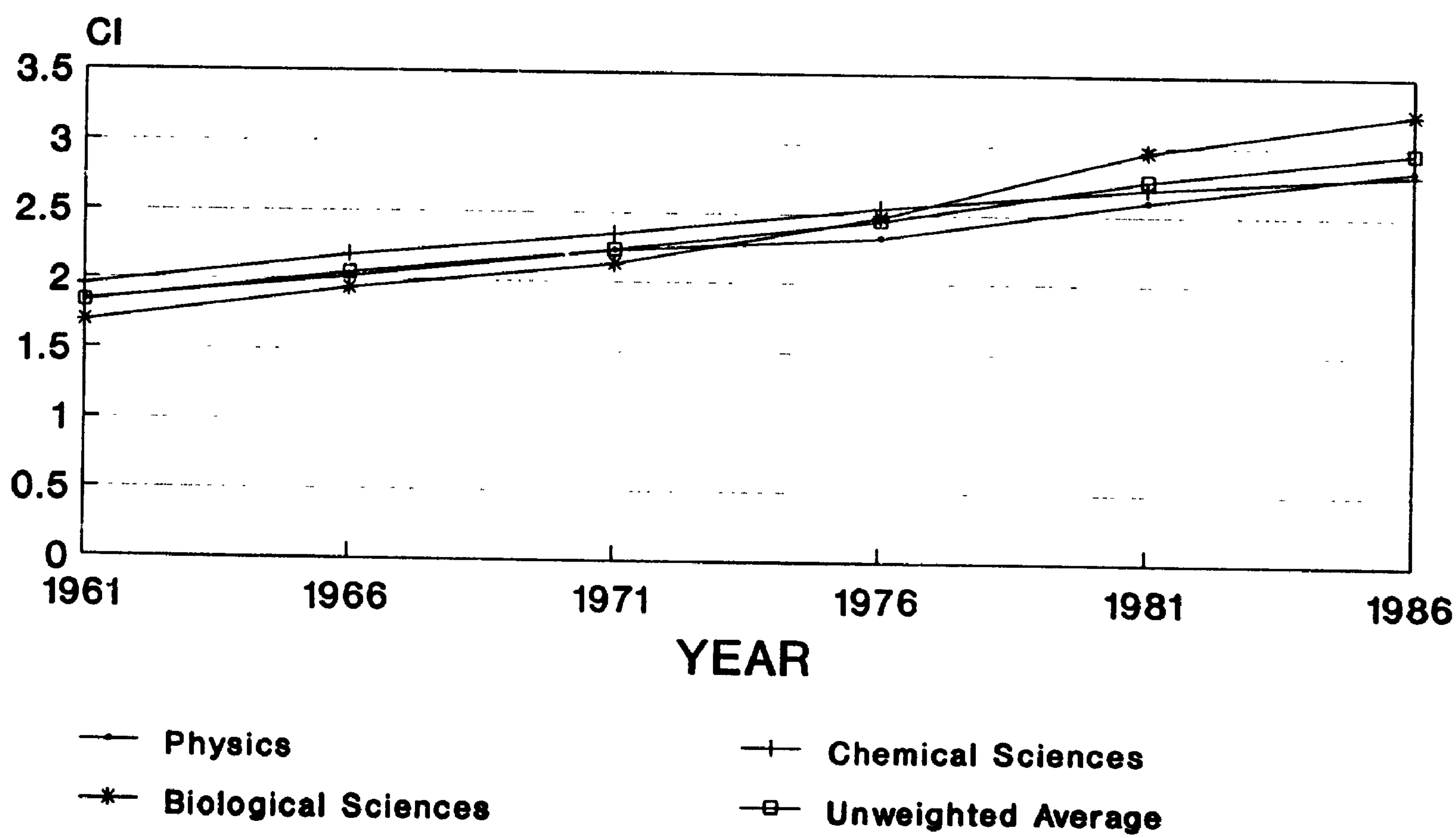


FIG. 4.4 : CI VERSUS YEAR
MATHEMATICAL SCIENCES

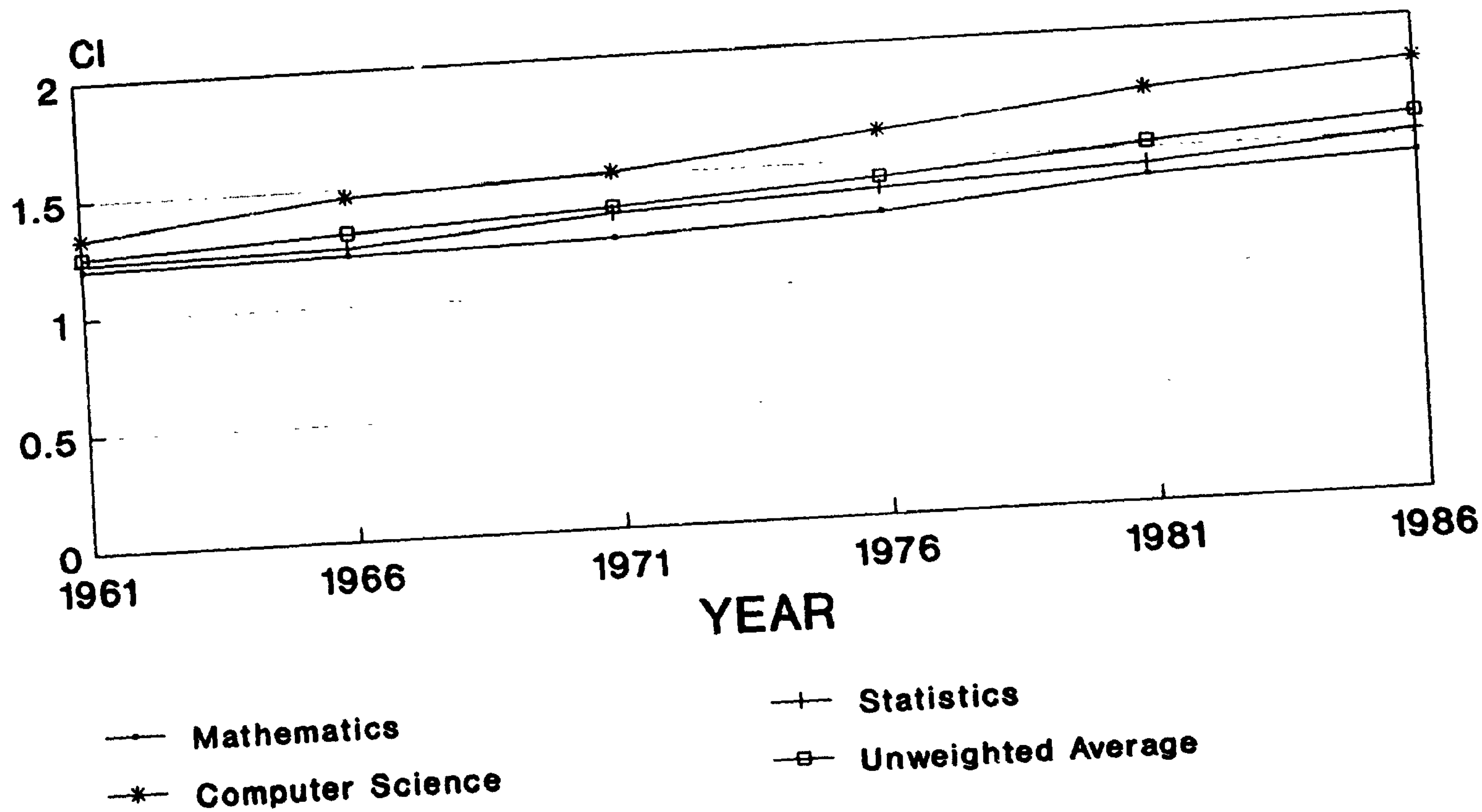


FIG. 4.5 : CI VERSUS YEAR
SOCIAL SCIENCES

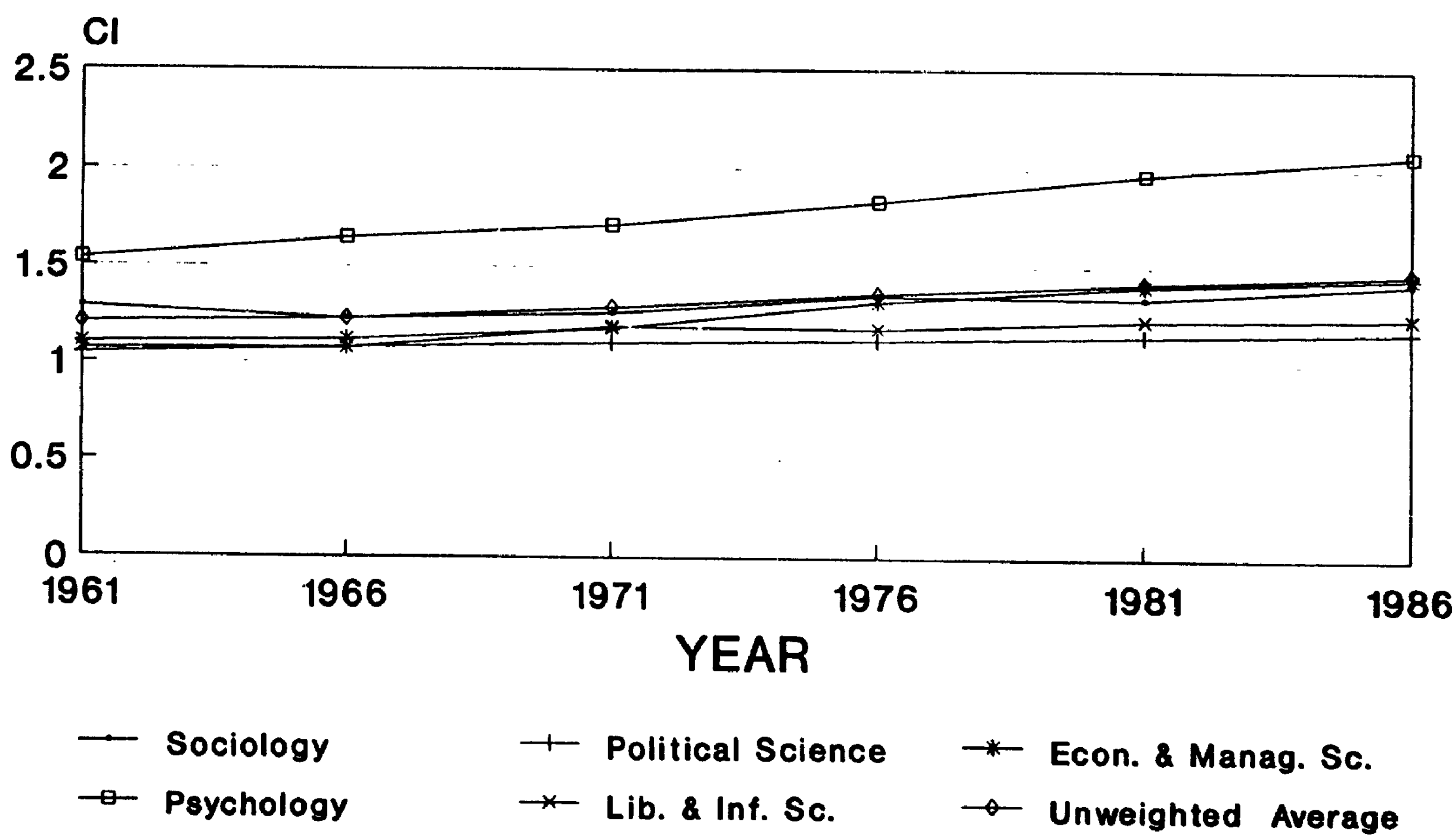
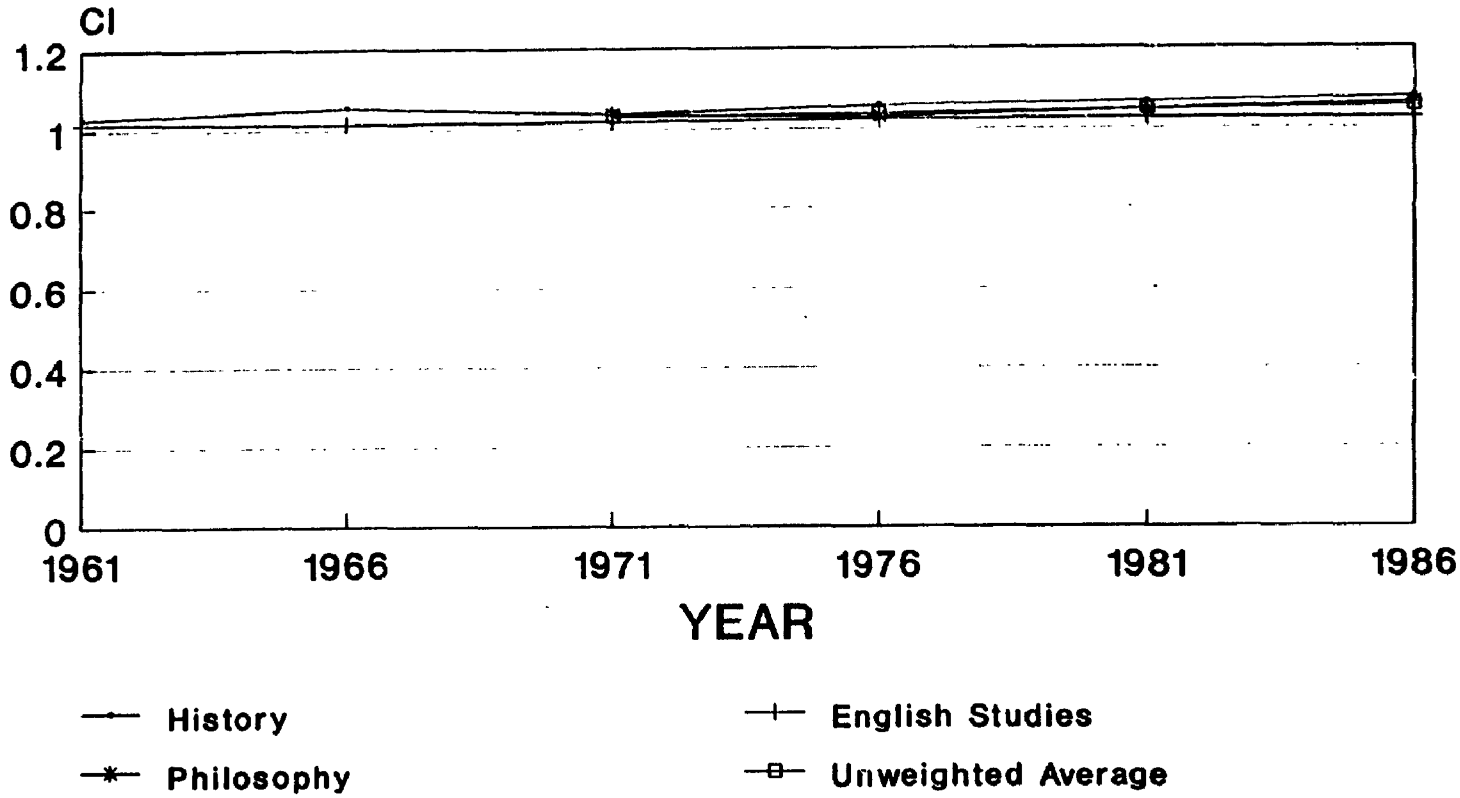


FIG. 4.6 : CI VERSUS YEAR
HUMANITIES



No data for Philosophy in 1961 and 1966

FIG. 4.7 : DC VERSUS YEAR
ENGINEERING SCIENCES

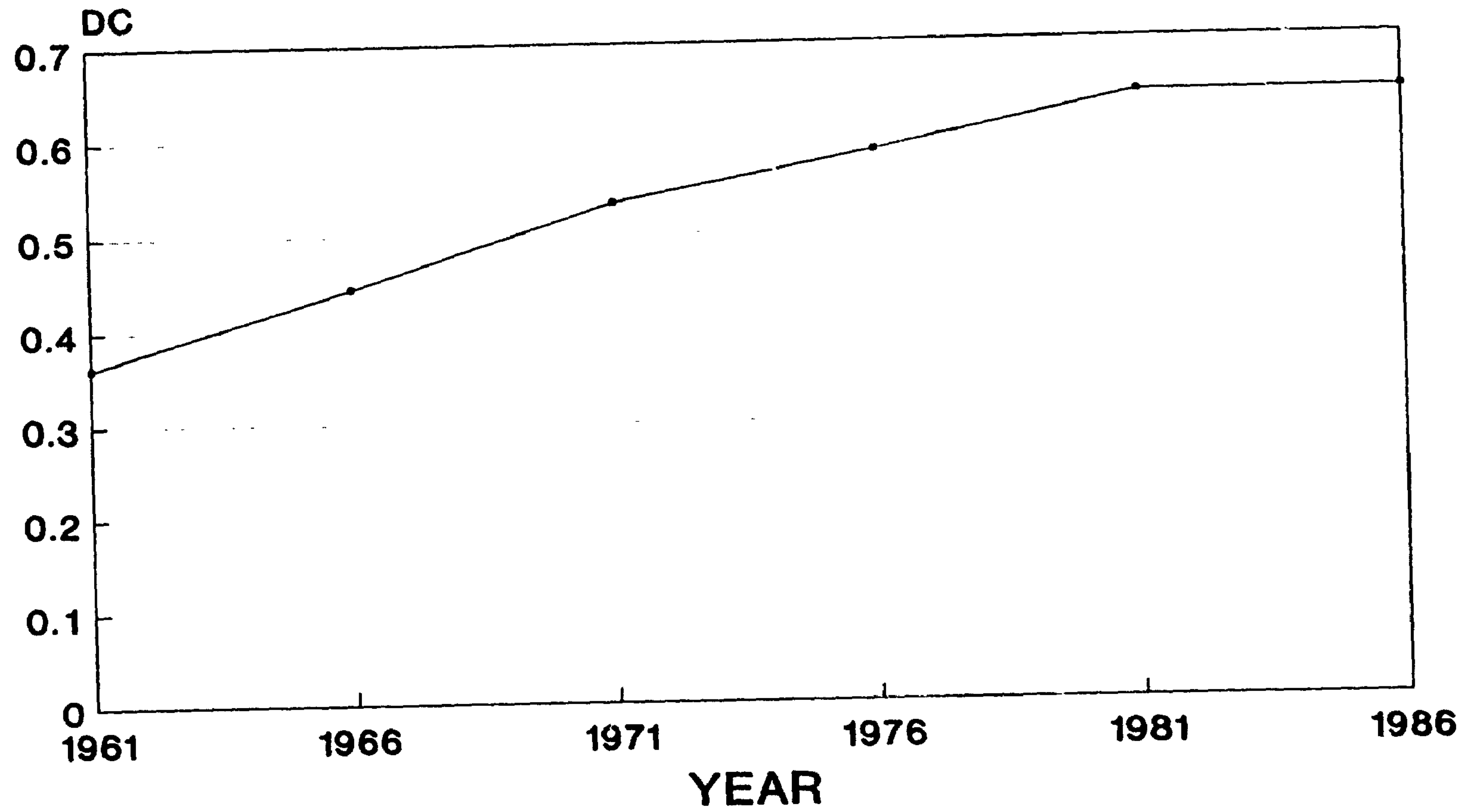


FIG. 4.8 : DC VERSUS YEAR
MEDICAL SCIENCES

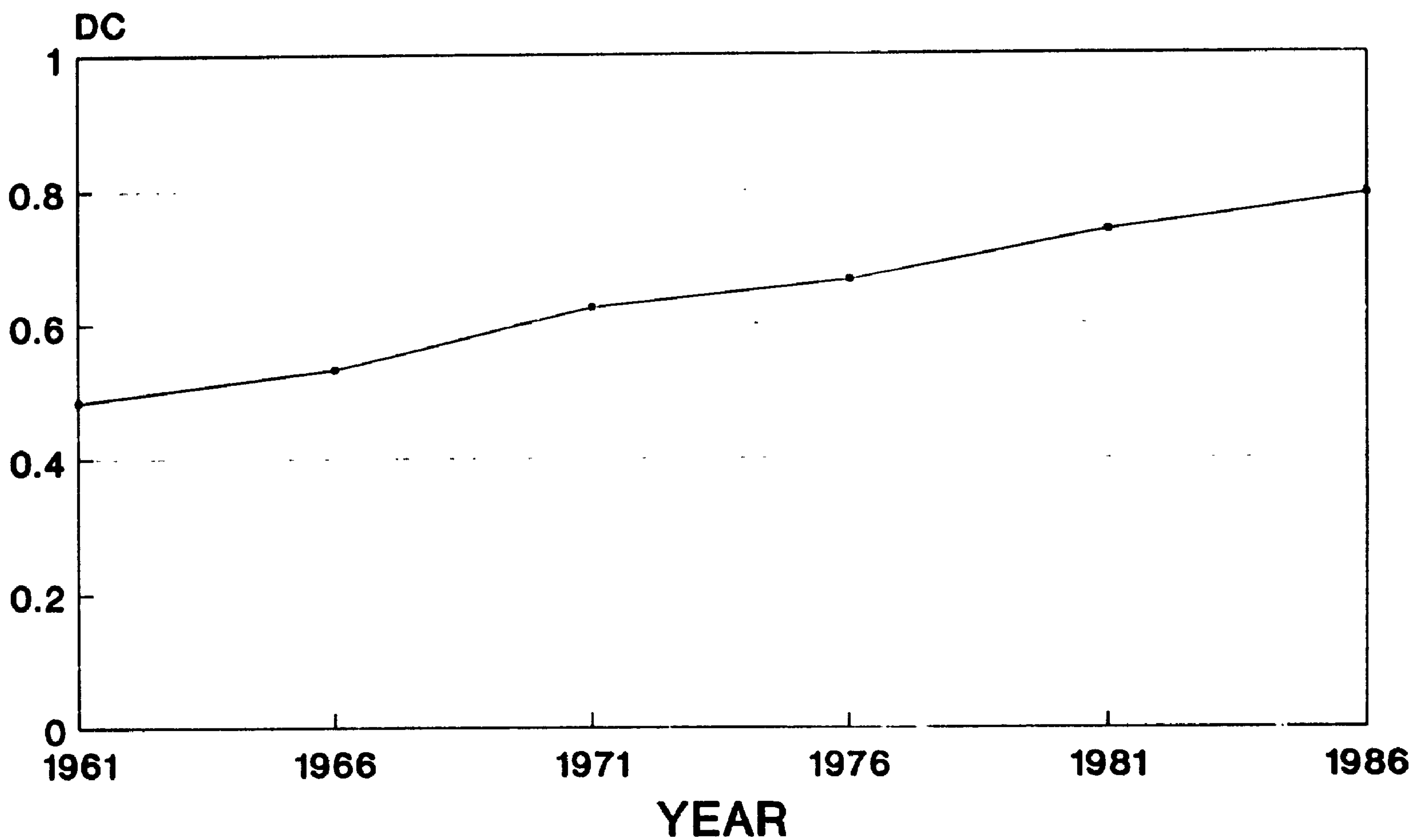


FIG. 4.9 : DC VERSUS YEAR
PHYSICAL SCIENCES

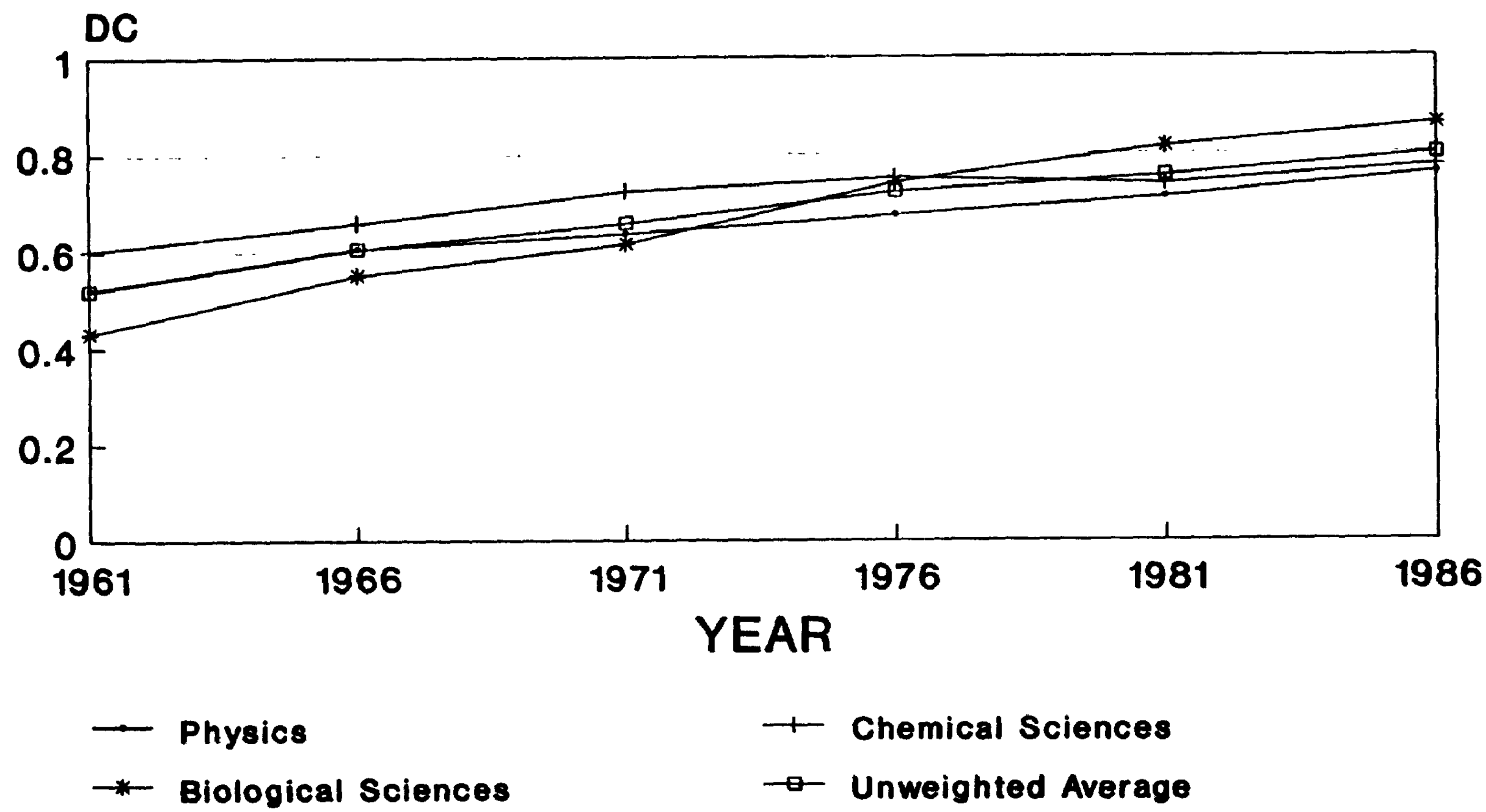


FIG. 4.10 : DC VERSUS YEAR
MATHEMATICAL SCIENCES

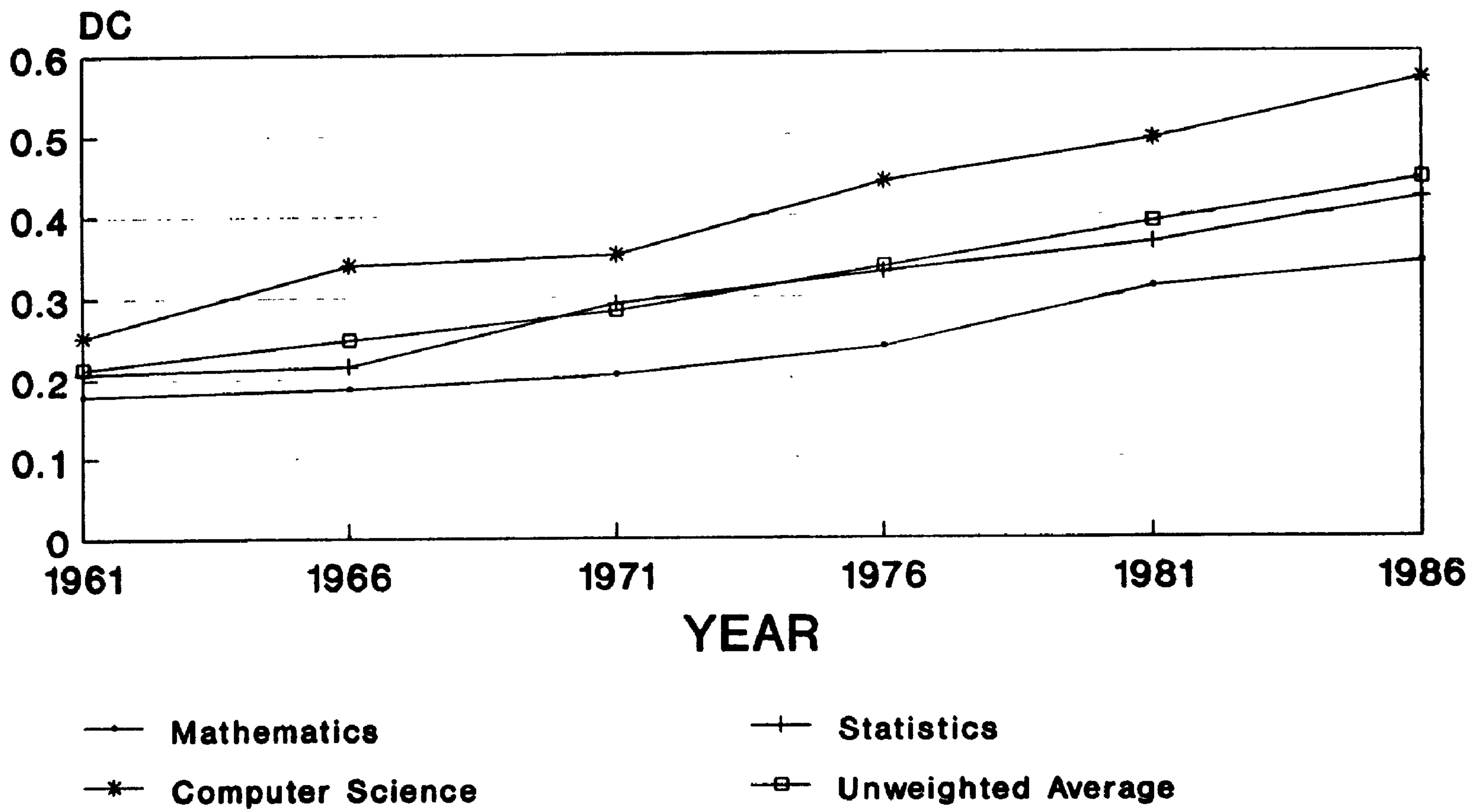


FIG. 4.11 : DC VERSUS YEAR

SOCIAL SCIENCES

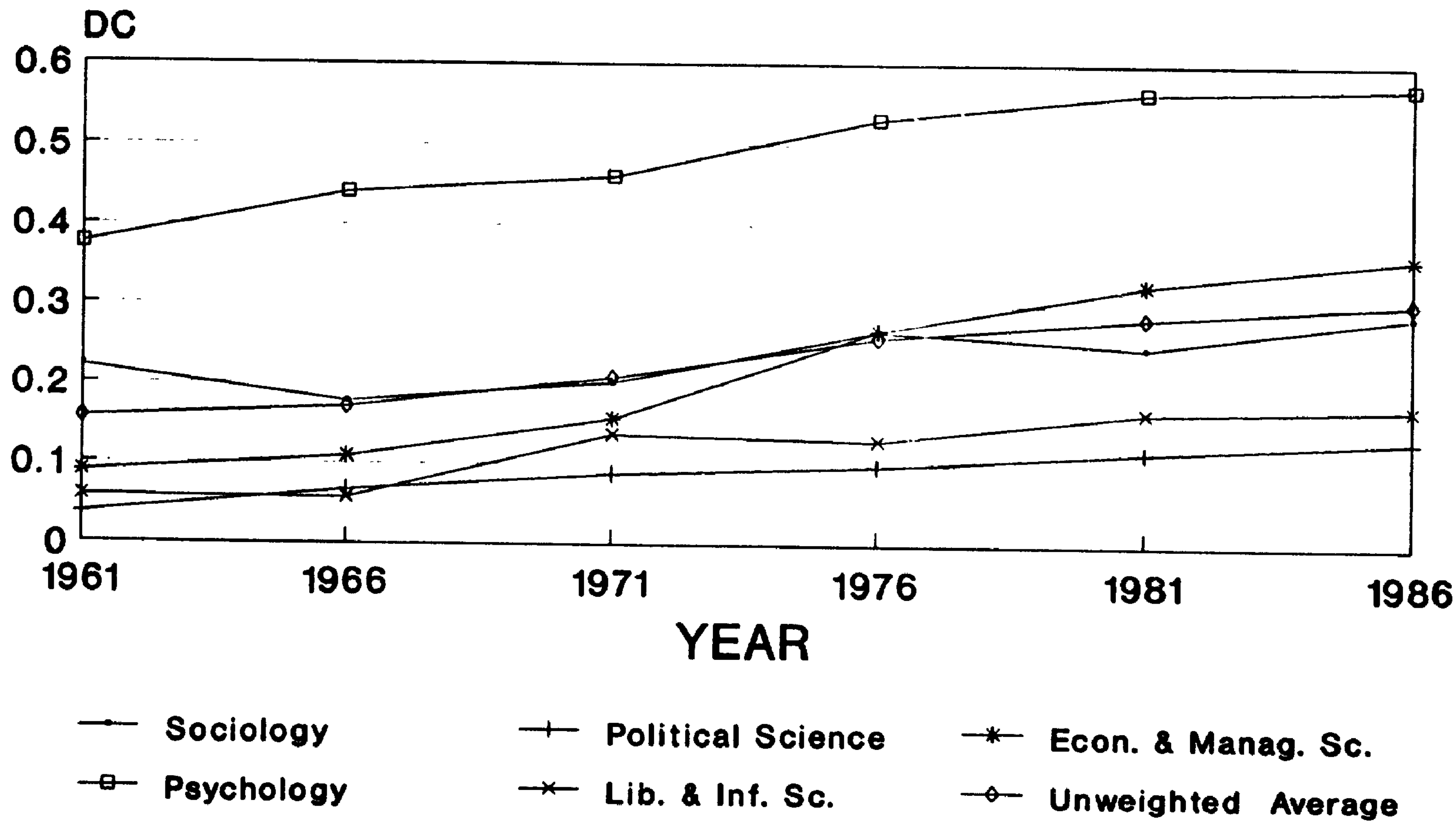
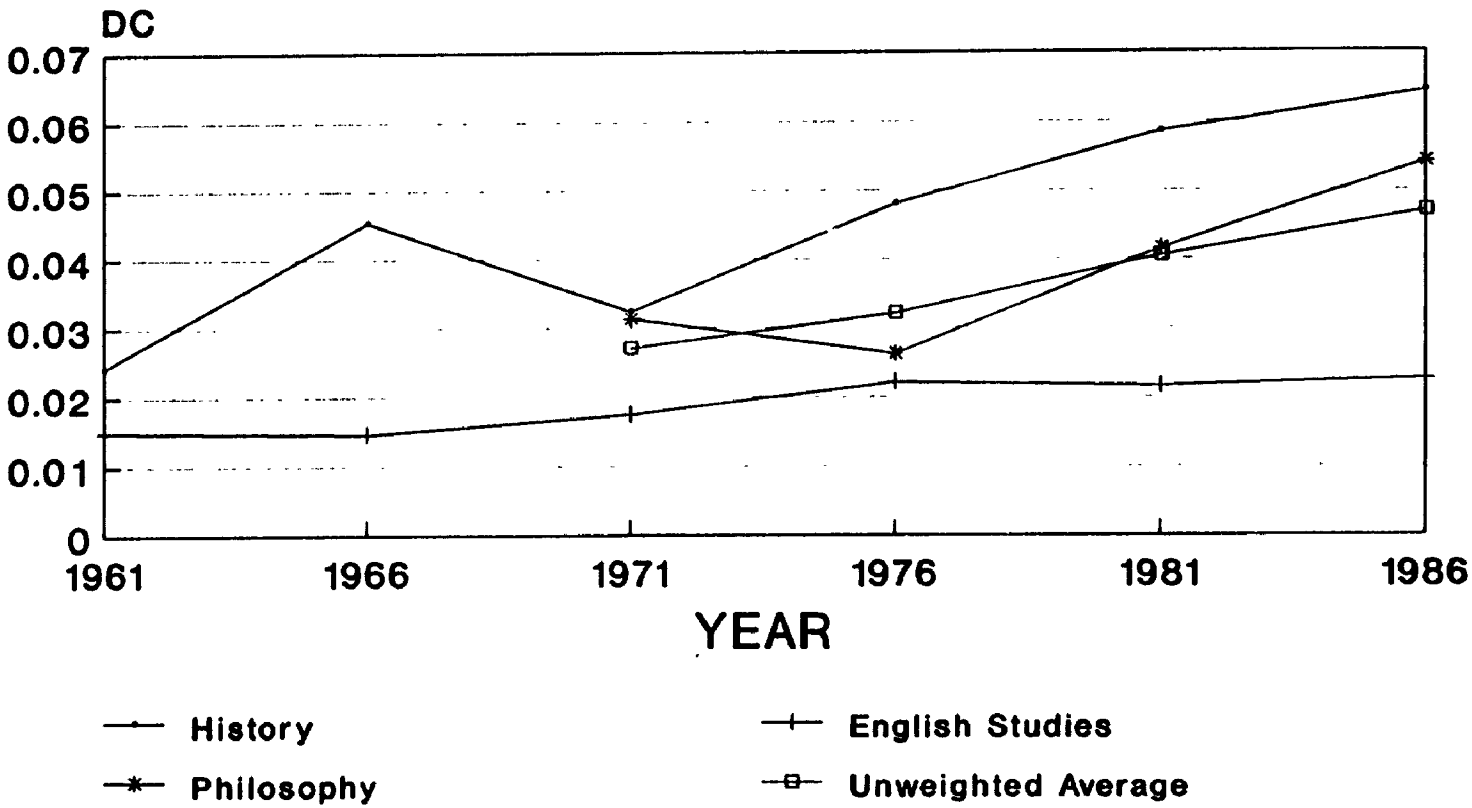


FIG. 4.12 : DC VERSUS YEAR
HUMANITIES



No data for Philosophy in 1961 and 1966

FIG. 4.13 : CC VERSUS YEAR
ENGINEERING SCIENCES

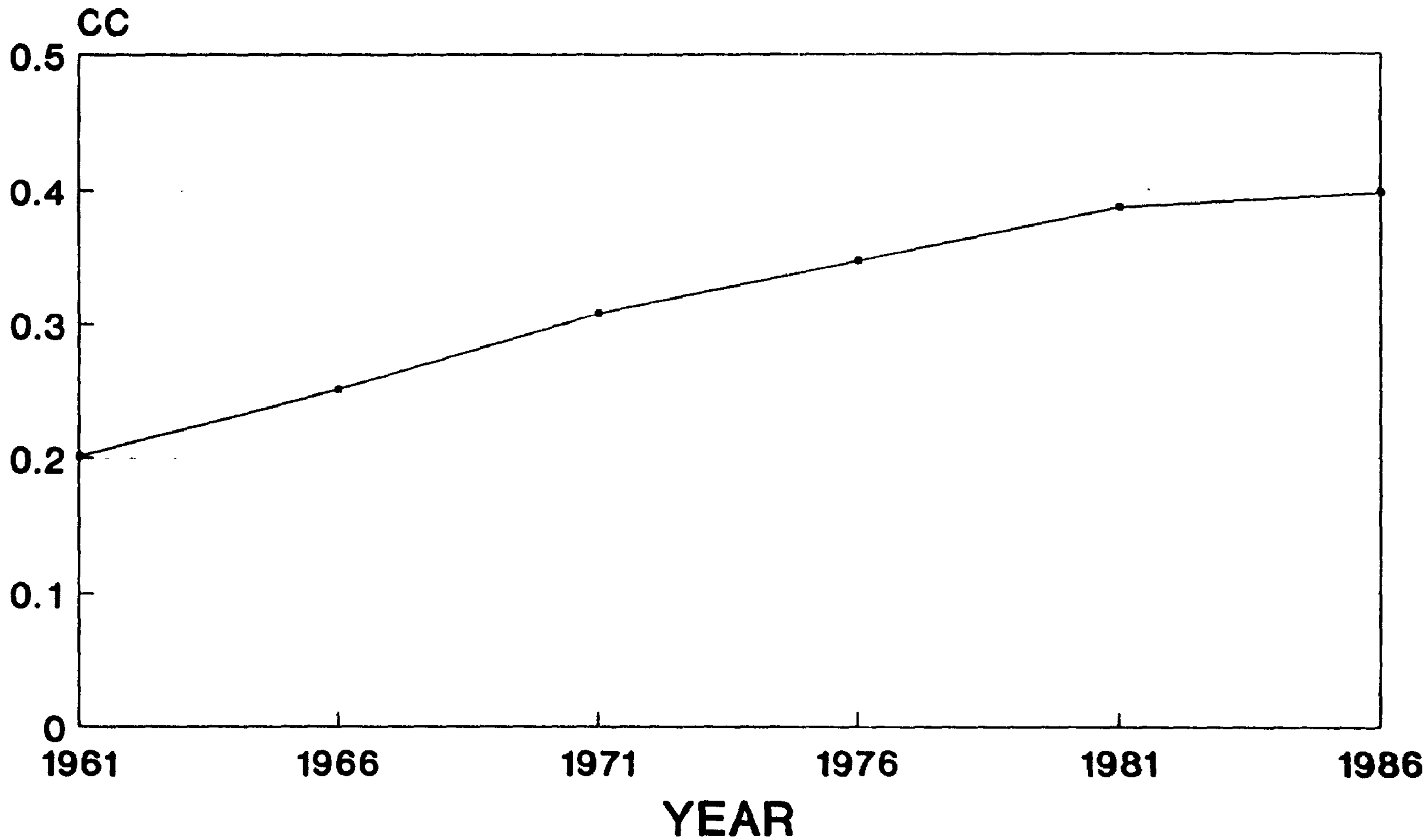


FIG. 4.14 : CC VERSUS YEAR
MEDICAL SCIENCES

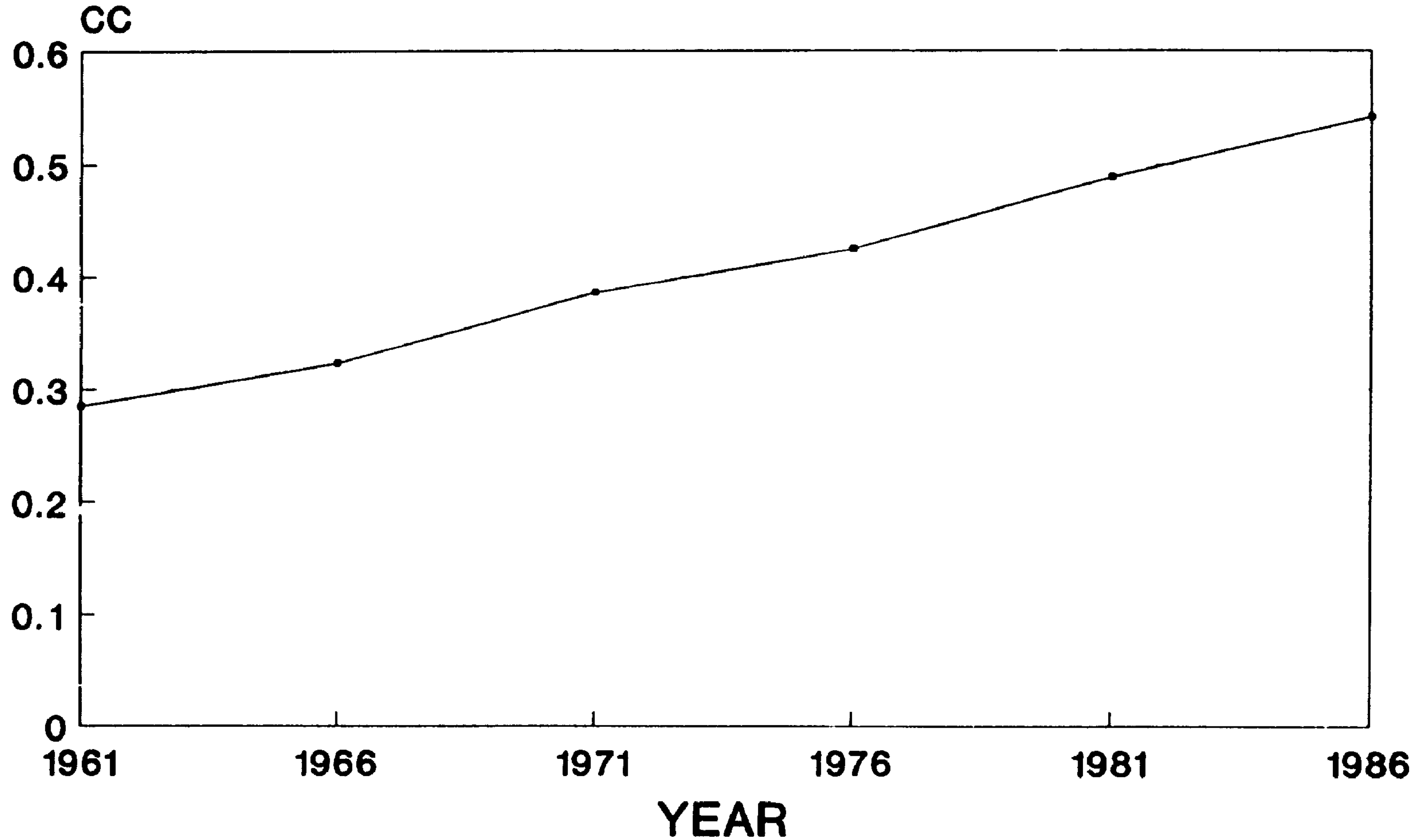


FIG. 4.15 : CC VERSUS YEAR
PHYSICAL SCIENCES

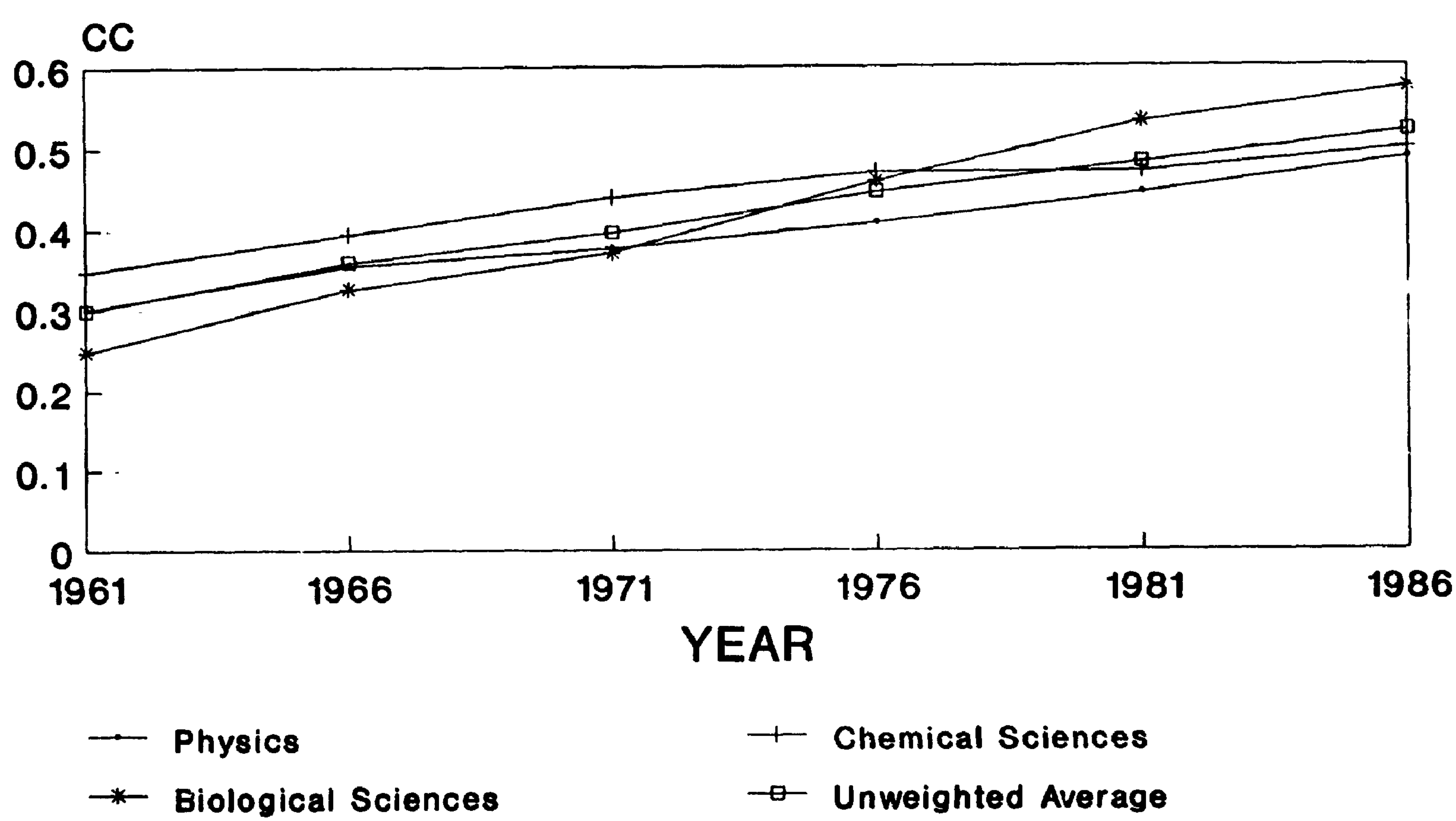


FIG. 4.16 : CC VERSUS YEAR
MATHEMATICAL SCIENCES

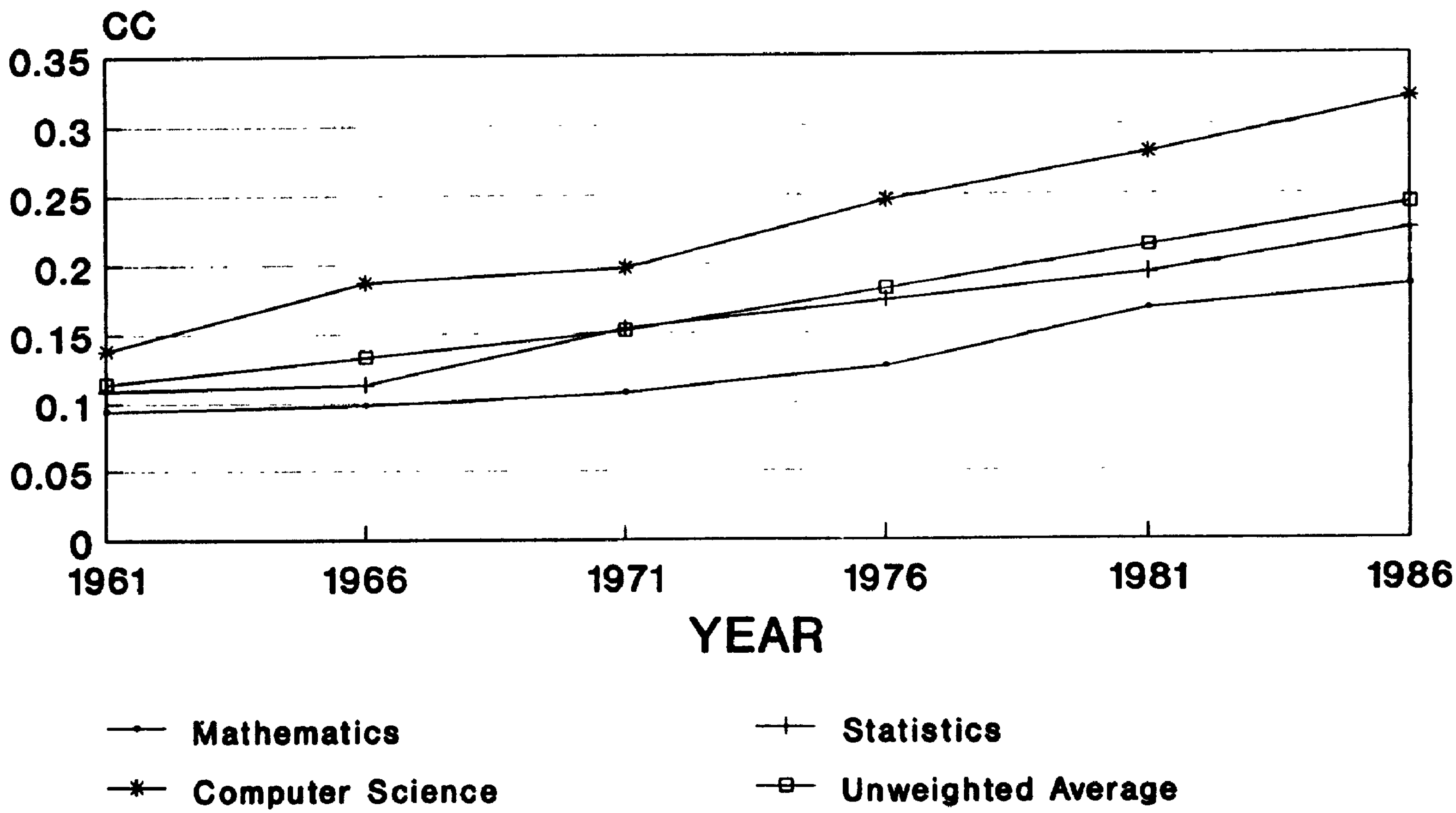


FIG. 4.17 : CC VERSUS YEAR
SOCIAL SCIENCES

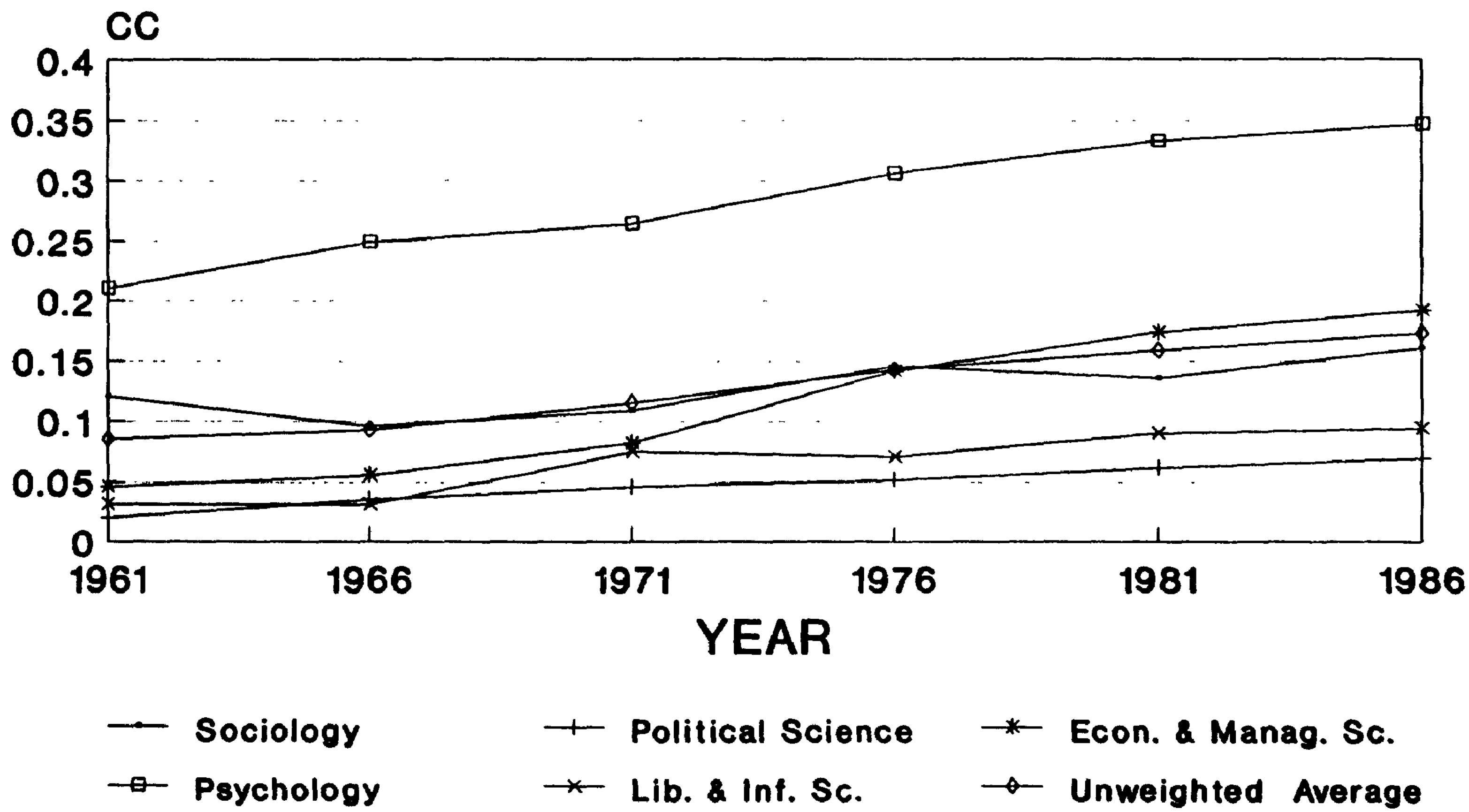
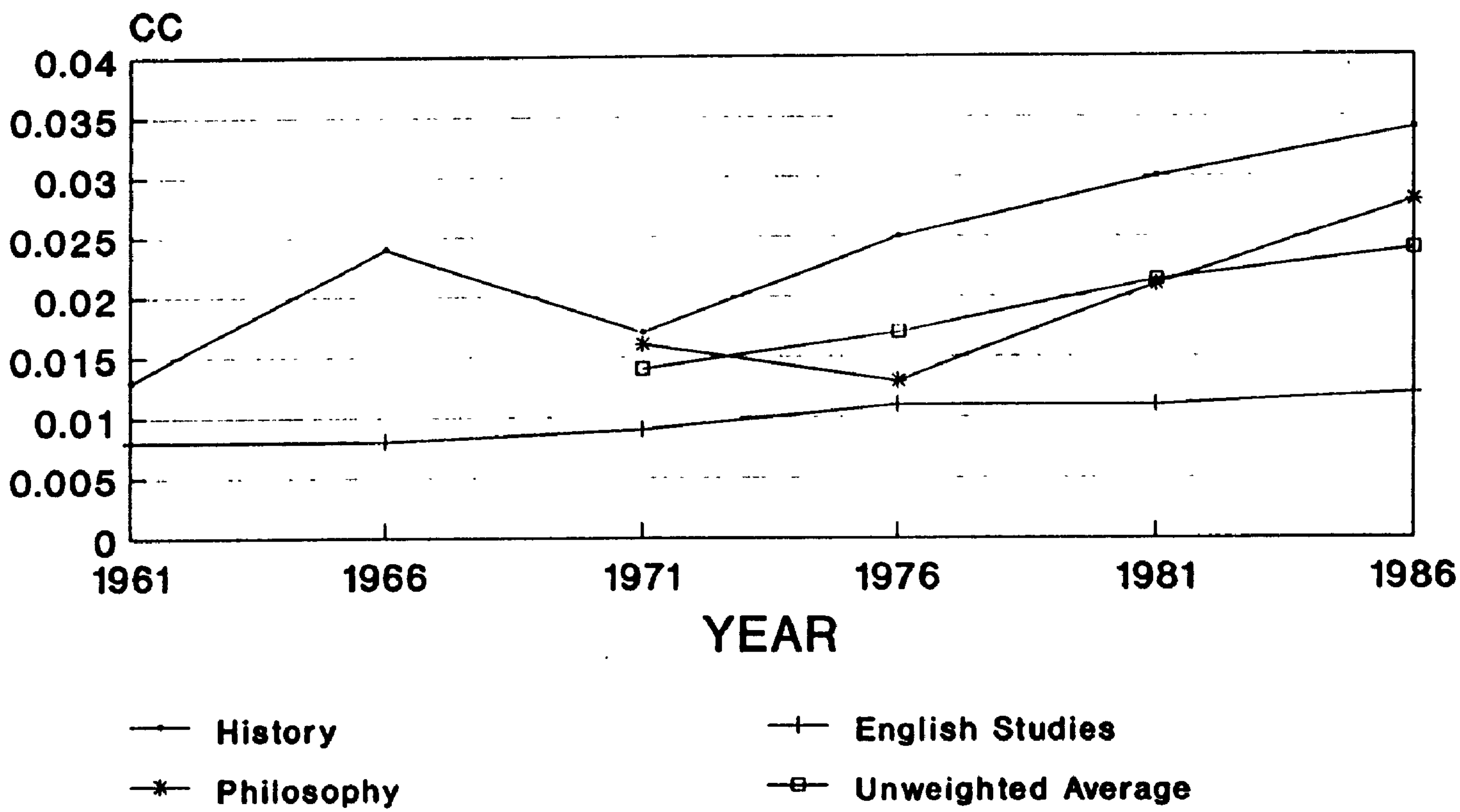


FIG. 4.18 : CC VERSUS YEAR
HUMANITIES



No data for Philosophy in 1961 and 1966

4.4 Comparisons of Collaboration in the Fields

CC was used to compare the degrees of collaboration in fields and decades. The hypotheses tested are :

(a) The degree of collaboration in all fields is the same

(b) The degree of collaboration in all decades is the same

(c) The effect of time on the degree of collaboration is the same for all fields

Because of unequal cell sizes (see the experimental design layout in Table 4.3 below), the two-way analysis of variance method for proportionate cell frequencies was used. The F-ratios for the three hypotheses are 102.18, 21.38 and 1.6 respectively (see Table 4.4). The first two are significant at the 1% level but the third one is not significant at the 5% level. Hence, we reject hypotheses (a) and (b) but accept hypothesis (c).

To find out which pairs of fields as well as pairs of decades differ, we employed the Scheffe's test (Berenson, Levine & Goldstein, 1983).

The results of the tests for the fields (see Table 4.5) show that there is less collaboration in the humanities than in other fields, less collaboration in the social sciences as well as in the mathematical sciences than in the engineering sciences, medical sciences and physical sciences.

The results of the tests for the decades (see Table 4.6) show that there was less collaboration in the '60s than in the two subsequent decades and there was less collaboration in the '70s than in the '80s.

Table 4.3 : Experimental Design Layout for Comparison of the Degrees of Collaboration in all the Fields

	Decades		
Fields	1960s	1970s	1980s
Engineering Sciences	$n_{11} = 2$	$n_{12} = 2$	$n_{13} = 2$
Medical Sciences	$n_{21} = 2$	$n_{22} = 2$	$n_{23} = 2$
Physical Sciences	$n_{31} = 6$	$n_{32} = 6$	$n_{33} = 6$
Mathematical Sciences	$n_{41} = 6$	$n_{42} = 6$	$n_{43} = 6$
Social Sciences	$n_{51} = 8$	$n_{52} = 8$	$n_{53} = 8$
Humanities	$n_{61} = 6$	$n_{62} = 6$	$n_{63} = 6$

Note : The within cell replication represents one of two different years within the decade for one of the abstracting journals within the field.

Table 4.4 : ANOVA Table for Tests of Equality of CC for Fields and Decades

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGN. OF F
MAIN EFFECTS	2.120	7	0.303	79.096	0.000
FIELD	1.956	5	0.391	102.182	0.000
DECADE	0.164	2	0.082	21.383	0.000
2-WAY INTERACTIONS	0.061	10	0.006	1.604	0.121
FIELD DECADE	0.061	10	0.006	1.604	0.121
EXPLAINED	2.181	17	0.128	33.513	0.000
RESIDUAL	0.299	78	0.004		
TOTAL	2.480	95	0.026		

Table 4.5 : Results of Scheffe's Multiple Comparison Tests for the Fields

Fields	Human.	Soc. Sc.	Math.Sc.	Eng.Sc.	Med.Sc.	Phy.Sc.
Humanit. (0.0172)						
Social Sciences (0.1282)	*					
Mathemat. Sciences (0.1727)	*					
Engineer. Sciences (0.3154)	*	*	*			
Medical Sciences (0.4081)	*	*	*			
Physical Sciences (0.4169)	*	*	*			

Note : 1) The average value of CC for each field is given in parentheses.

2) * indicates that the CC value for the column field is significantly less than that of the row field.

Table 4.6 : Results of Scheffe's Multiple Comparison Tests for the Decades

Decades	1960s	1970s	1980s
1960s (0.1484)			
1970s (0.1993)	*		
1980s (0.2495)	*	*	

Note : 1) The average value of CC for each decade is given in parentheses.

2) * indicates that the CC value for the column decade is significantly less than that of the row decade.

CHAPTER 5

Models for the Distribution of Authorships

The number of authors of a paper is often taken as an indicator of collaboration in research. In real situations, this variable can obviously take only discrete values. It is therefore reasonable to assume that a discrete probability model will describe the distribution of authorships. In this chapter, a theoretical model for the distribution of authorships is derived. This model and other common discrete probability models are tested for their goodness-of-fit. Since the distribution of authorships has no value for the zero case (there is no paper without an author), we use the shifted form of any model with a domain starting from zero. A shifted discrete distribution is one in which the independent variable x is transformed to x+1, and the values now start from one.

5.1 Theoretical Model

The initial theoretical model being proposed is the shifted Waring distribution. Its derivation is as follows :

Assume that a researcher completes a project alone if he or she can. However, if he or she cannot, he or she brings in additional researchers, one at a time, until the project is completed. Let an attempt by the researcher(s) to complete a project without bringing in an additional researcher represent a trial. The probability that the project will be completed on any trial is constant. The distribution of the number of trials or the number of researchers required before the project is completed can then be described by the geometric distribution. However, the probability of completing the project at any trial without bringing in an additional researcher, i.e., p, would vary from project to project (types of projects include review, opinion on a topic, experimental research, theoretical work, etc). Thus,

$$P(X = r \mid p) = p\,(1-p)^{r-1} \quad ; r = 1,2,\ldots$$

For analytical convenience, we assume that p varies as a beta distribution with parameters α,β, i.e., p has a density function

$$f(p) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\, p^{\alpha-1}\,(1-p)^{\beta-1} & \text{if } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 0$, $\beta > 0$ are constants and

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}\,(1-x)^{\beta-1}\,dx$$

$$= \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Hence,

$$P(X = r) = E\,[P(X = r \mid p)]$$

$$= \int P(X = r \mid p)\, f(p)\, dp$$

$$= \int_0^1 p(1-p)^{r-1} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \, dp$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 p^{\alpha} (1-p)^{r+\beta-2} \, dp$$

$$= \frac{1}{B(\alpha,\beta)} B(\alpha+1, r+\beta-1)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \frac{\Gamma(\alpha+1)\,\Gamma(r+\beta-1)}{\Gamma(\alpha+\beta+r)}$$

$$= \frac{\alpha}{\alpha+\beta} \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+\beta+r)} \frac{\Gamma(r+\beta-1)}{\Gamma(\beta)}$$

$$= \frac{\alpha}{\alpha+\beta} \prod_{k=1}^{r} \frac{(k+\beta-2)}{(\alpha+\beta+k-1)} \quad ; \ \alpha > 0, \ \beta > 0, \ r = 1,2,\ldots$$

This distribution is known as the shifted Waring distribution.

5.2 Other Discrete Probability Models

For comparative purposes, other discrete probability models that have been suggested for other bibliometric phenomena are considered. These are :

1. Zipf
2. Mandelbrot
3. Geometric
4. Shifted Poisson
5. Shifted Generalized Poisson
6. Logarithmic

7. Borel-Tanner
8. Shifted Yule
9. Shifted Generalized Waring
10. Shifted Inverse Gaussian-Poisson (IGP)
11. Shifted Generalized Inverse Gaussian-Poisson (GIGP)
12. Shifted Binomial
13. Shifted Beta-Binomial
14. Shifted Negative Binomial
15. Shifted Generalized Negative Binomial

The form of each model including the shifted Waring and the method of estimation of its parameters are given in Appendix C. More details about these models can be found in Mandelbrot (1953), Johnson and Kotz (1969), Chatfield and Goodhardt (1970), Jain and Consul (1971), Consul and Jain (1973), Irwin (1975), Consul and Shoukri (1984), and Sichel (1985).

5.3 Goodness-of-fit Tests

The Chi Square test was used to determine the goodness-of-fit of the models for the distribution of authorships. The Chi Square test was chosen in preference to the Kolmogorov-Smirnov test because the latter is strictly applicable only if the data are continuously distributed

(Conover, 1980). Computer programs written in Basic and Pascal were used to perform the tests.

The result of the test was considered invalid if either :

(a) the parameters of a model could not be computed because of computational difficulties or non-existence of either the maximum likelihood or moment estimators ; or

(b) the assumptions of the model are violated.

A model is considered to have passed the test if there is no significant difference between the observed and the expecred distributions at the specified level of significane while a model is considered to have provided the best fit to a data set if its Chi Square value has the highest probability level. A model is also considered to overfit a data set if the degree of freedom for the Chi Square value is less than one.

The summary results of the Chi Square tests for each field are given below (the results for the disciplines within the fields are given in Appendix D) :

(i) Humanities

Many of the models performed very well with Mandelbrot, shifted generalized Poisson, shifted Waring, shifted Inverse Gaussian-Poisson and shifted negative binomial distributions doing exceptionally well (see Table 5.1 below). It is noted that some of the models overfit some of the data sets in this field.

Table 5.1 : Results of the Chi Square Goodness-of-fit Test for Humanities (18 data sets)

Model	No. of Valid Results	No. of Passes 5%	1%	No. of Best Fits Provided
Zipf	18	9	12	1
Mandelbrot	18	14	16	4
Geometric	18	6	10	0
Shifted Poisson	18	7	7	1
Sh. Gen. Poisson	18	17	17	1
Logarithmic	18	7	7	0
Borel-Tanner	18	7	9	0
Shifted Yule	18	10	13	3
Shifted Waring	17	14	17	0
Sh. Gen. Waring	3	0	0	0
Shifted IGP	17	16	17	6
Shifted GIGP	2	2	2	0
Shifted Binomial	1	1	1	0
Sh. Beta-Binomial	17	0	0	0
Sh. Neg. Binomial	17	14	16	0
Sh. Gen. Neg. Bin.	14	13	14	2

Table 5.2 : Results of the Chi Square Goodness-of-fit Test for Social Sciences (30 data sets)

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	30	4	4	0
Mandelbrot	30	9	11	1
Geometric	30	6	8	0
Shifted Poisson	30	3	4	1
Sh. Gen. Poisson	30	23	27	2
Logarithmic	30	7	9	0
Borel-Tanner	30	8	10	0
Shifted Yule	30	8	10	2
Shifted Waring	18	16	18	2
Sh. Gen. Waring	17	1	2	0
Shifted IGP	30	25	30	16
Shifted GIGP	8	6	7	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	30	4	10	0
Sh. Neg. Binomial	30	21	25	4
Sh. Gen. Neg. Bin.	19	14	17	2

Table 5.3 : Results of the Chi Square Goodness-of-fit Test for Mathematical Sciences (18 data sets)

Model	No. of Valid Results	No. of Passes 5%	1%	No. of Best Fits Provided
Zipf	18	0	0	0
Mandelbrot	18	0	0	0
Geometric	18	4	5	1
Shifted Poisson	18	2	4	0
Sh. Gen. Poisson	18	7	9	2
Logarithmic	18	0	0	0
Borel-Tanner	18	0	0	0
Shifted Yule	18	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	4	0	0	0
Shifted IGP	14	5	6	13
Shifted GIGP	1	0	1	0
Shifted Binomial	4	3	3	0
Sh. Beta-Binomial	6	3	3	0
Sh. Neg. Binomial	14	3	5	0
Sh. Gen. Neg. Bin.	5	4	5	2

Table 5.4 : Results of the Chi Square Goodness-of-fit Test for Physical Sciences (18 data sets)

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	18	0	0	0
Mandelbrot	18	0	0	0
Geometric	18	0	0	0
Shifted Poisson	18	0	0	0
Sh. Gen. Poisson	18	1	2	2
Logarithmic	18	0	0	0
Borel-Tanner	18	0	0	0
Shifted Yule	18	0	0	0
Shifted Waring	2	0	0	0
Sh. Gen. Waring	18	0	0	0
Shifted IGP	18	3	5	15
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	7	0	0	0
Sh. Neg. Binomial	18	1	1	1
Sh. Gen. Neg. Bin.	1	0	0	0

(ii) Social Sciences

Only shifted generalized Poisson, shifted inverse Gaussian-Poisson and shifted negative binomial distributions did very well while shifted Waring and shifted generalized negative binomial distributions each fitted about 60% of the data sets (see Table 5.2 above). A few cases of overfitting by some of the models also occurred in this field.

(iii) Mathematical Sciences

Most of the models performed poorly but the shifted generalized Poisson fitted about half of the data sets while the shifted inverse Gaussian-Poisson fitted about one-third of the data sets (see Table 5.3 above). However, the shifted inverse Gaussian-Poisson provided the best fit to about 72% of the data sets.

(iv) Physical Sciences

All the models performed poorly. However, the shifted inverse Gaussian-Poisson provided the best fit to about 83% of the data sets (see Table 5.4 above). It seems that as collaboration increases, the further the expected values depart from the observed values. It is also noted that in this field, the mode of most of the observed distributions is 2.

(v) Medical Sciences

All the models performed badly with only shifted generalized Poisson, shifted inverse Gaussian-Poisson, shifted beta-binomial, shifted negative binomial and shifted generalized negative binomial distributions fitting one or two data sets (see Table 5.5 below). The five models only fitted data sets in the '60s but as collaboration increased rapidly in the '70s, the expected values started to depart significantly from the observed values.

(vi) Engineering Sciences

All the models performed badly here too with shifted generalized Poisson, shifted inverse Gaussian-Poisson and shifted negative binomial distributions fitting only the 1961 data set (see Table 5.6 below). However, the shifted inverse Gaussian-Poisson provided the best fit to all the data sets.

Overall, only shifted generalized Poisson, shifted Waring, shifted inverse Gaussian-Poisson, shifted negative binomial and shifted generalized negative binomial distributions performed fairly well (see Table 5.7 below). However, the moment or maximum likelihood estimators didn't exist for about 61% of the data sets in the case of the shifted Waring distribution, which is also the proposed theoretical model, while for about 55% of the data sets, no real roots existed for one of the parameters of the shifted

Table 5.5 : Results of the Chi Square Goodness-of-fit Test for Medical Sciences (6 data sets)

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	1	2	0
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	4	0	0	0
Shifted IGP	6	1	1	1
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	0	1	0
Sh. Neg. Binomial	6	2	2	3
Sh. Gen. Neg. Bin.	4	2	2	2

Table 5.6 : Results of the Chi Square Goodness-of-fit Test for Engineering Sciences (6 data sets)

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	1	1	0
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	6	0	1	0
Shifted IGP	6	1	1	6
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	4	0	0	0
Sh. Neg. Binomial	6	1	1	0
Sh. Gen. Neg. Bin.	0	0	0	0

Table 5.7 : Overall Results of the Chi Square Goodness-of-fit Test (96 data sets)

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	96	13	16	1
Mandelbrot	96	23	27	5
Geometric	96	16	23	1
Shifted Poisson	96	12	15	2
Sh. Gen. Poisson	96	50	58	7
Logarithmic	96	14	16	0
Borel-Tanner	96	15	19	0
Shifted Yule	96	18	23	5
Shifted Waring	37	30	35	2
Sh. Gen. Waring	52	1	3	0
Shifted IGP	91	51	60	57
Shifted GIGP	11	8	10	0
Shifted Binomial	5	4	4	0
Sh. Beta-Binomial	70	7	14	0
Sh. Neg. Binomial	91	42	50	8
Sh. Gen. Neg. Bin.	43	33	38	8

generalized negative binomial. As for the other three models, the shifted generalized Poisson fitted about 60% of the data sets and provided the best fit to about 7%, shifted inverse Gaussian-Poisson fitted about 63% and provided the best fit to about 59% while the shifted negative binomial fitted about 52% and provided the best fit to about 8% of the data sets. There is no significant difference between the proportions of times the three models passed the Chi Square tests but the proportion of best fits provided by snifted inverse Gaussian-Poisson is significantly greater than the proportion provided by either shifted negative binomial or shifted generalized Poisson. Hence, on the basis of the number of best fits provided, the shifted inverse Gaussian-Poisson is chosen over the other two models as the probability model for the distribution of authorships (see Appendix E for the fits of the shifted inverse Gaussian-Poisson to the data sets for 1986). One would have expected the shifted generalized inverse Gaussian-Poisson to perform better than the shifted inverse Gaussian-Poisson since the latter is a special case of the former, but we couldn't estimate the parameter τ for about 88% of the data sets. The shifted inverse Gaussian-Poisson is obtained from the shifted generalized inverse Gaussian-Poisson by setting the value of τ to $-\frac{1}{2}$. The derivation of the shifted inverse Gaussian-Poisson as a model for the distribution of authorships is given in Appendix F.

5.4 Estimates of the Parameters of the Shifted Inverse Gaussian-Poisson Distribution

The estimates of the parameters of the shifted inverse Gaussian-Poisson for all the data sets are tabulated in Table 5.8 and plotted in Figures 5.1 to 5.12. It is observed that :

(i) Engineering Sciences

α ranges between 2 and 4 while θ ranges between 0.3 and 0.6. Both parameters tend to increase with time.

(ii) Medical Sciences

α ranges between 2 and 5 while θ ranges between 0.4 and 0.7. Both parameters tend to increase with time.

(iii) Physical Sciences

α ranges between 1 and 7 while θ ranges between 0.2 and 0.7. However, both parameters tend to fluctuate with time.

(iv) Mathematical Sciences

The estimates of both parameters for statistics data sets in 1971, 1976, 1981 and 1986 are negative. These values violate the assumptions of the model and are, therefore, considered invalid. For the other two disciplines in this field, α ranges between 1 and 12 while θ ranges between 0.09 and 0.5. Both parameters tend to fluctuate with time.

(v) Social Sciences

α values for different disciplines vary considerably. For two of the disciplines, political science and library and information science, α is less than one but tends to increase with time. For sociology, the value lies in the neighbourhood of one. The value for psychology ranges between 2 and 4.5 but tends to fluctuate with time. The value for economics ranges between 0.9 and 11 and tends to increase with time.

The values of θ for the various disciplines seem more homogeneous than the α values. For four of the disciplines, the values range between 0.25 and 0.6 but they tend to fluctuate with time. For the fifth discipline, economics, the value ranges between 0.07 and 0.19 but tends to decrease with time.

(vi) Humanities

α ranges between 0.08 and 0.45 and tends to increase with time. θ ranges between 0.16 and 0.5 but tends to fluctuate with time.

Table 5.8 : Estimates of the Parameters of the Shifted Inverse Gaussian-Poisson Distribution

Field		Year 1961	1966	1971	1976	1981	1986
Engineering Sc.	α	2.06	2.35	3.11	3.41	3.64	3.35
	θ	0.39	0.43	0.42	0.45	0.47	0.52
Medical Sc.	α	2.53	2.40	3.05	3.23	4.32	4.47
	θ	0.46	0.55	0.55	0.58	0.56	0.62
Physical Sc.							
Physics	α	2.53	4.35	3.25	2.71	2.78	3.08
	θ	0.48	0.37	0.49	0.61	0.66	0.67
Chemical Sc.	α	5.91	4.17	6.95	6.61	2.63	3.23
	θ	0.27	0.43	0.32	0.37	0.70	0.66
Biological Sc.	α	1.72	3.35	3.45	6.60	5.44	5.97
	θ	0.54	0.43	0.47	0.36	0.51	0.52
Mathematical Sc.							
Mathematics	α	1.63	1.68	4.71	2.94	4.80	4.59
	θ	0.23	0.23	0.09	0.17	0.15	0.17
Statistics	α	5.99	2.75	*	*	*	*
	θ	0.08	0.17	*	*	*	*
Computer Sc.	α	1.29	2.05	1.60	3.15	2.99	11.16
	θ	0.40	0.36	0.47	0.33	0.39	0.14
Social Sc.							
Sociology	α	1.01	0.79	1.00	1.58	1.10	1.11
	θ	0.43	0.43	0.40	0.36	0.45	0.52
Political Sc.	α	0.17	0.37	0.48	0.44	0.60	0.60
	θ	0.39	0.34	0.34	0.41	0.37	0.41
Economics	α	0.95	1.97	3.01	7.03	10.57	4.39
	θ	0.19	0.11	0.11	0.09	0.07	0.18
Psychology	α	2.29	3.17	2.87	4.29	3.10	2.76
	θ	0.37	0.33	0.39	0.32	0.46	0.53
Lib.& Inf. Sc.	α	0.40	0.17	0.41	0.38	0.57	0.56
	θ	0.28	0.58	0.59	0.59	0.53	0.56
Humanities							
History	α	0.09	0.17	0.19	0.24	0.44	0.30
	θ	0.49	0.47	0.32	0.36	0.26	0.39
Engl. Studies	α	0.10	0.08	0.21	0.27	0.15	0.23
	θ	0.28	0.33	0.16	0.16	0.27	0.19
Philosophy	α	**	**	0.22	0.31	0.31	0.41
	θ	**	**	0.27	0.16	0.25	0.25

* Invalid estimates for the year

** No data for the year

FIG. 5.1 : α VERSUS YEAR
ENGINEERING SCIENCES

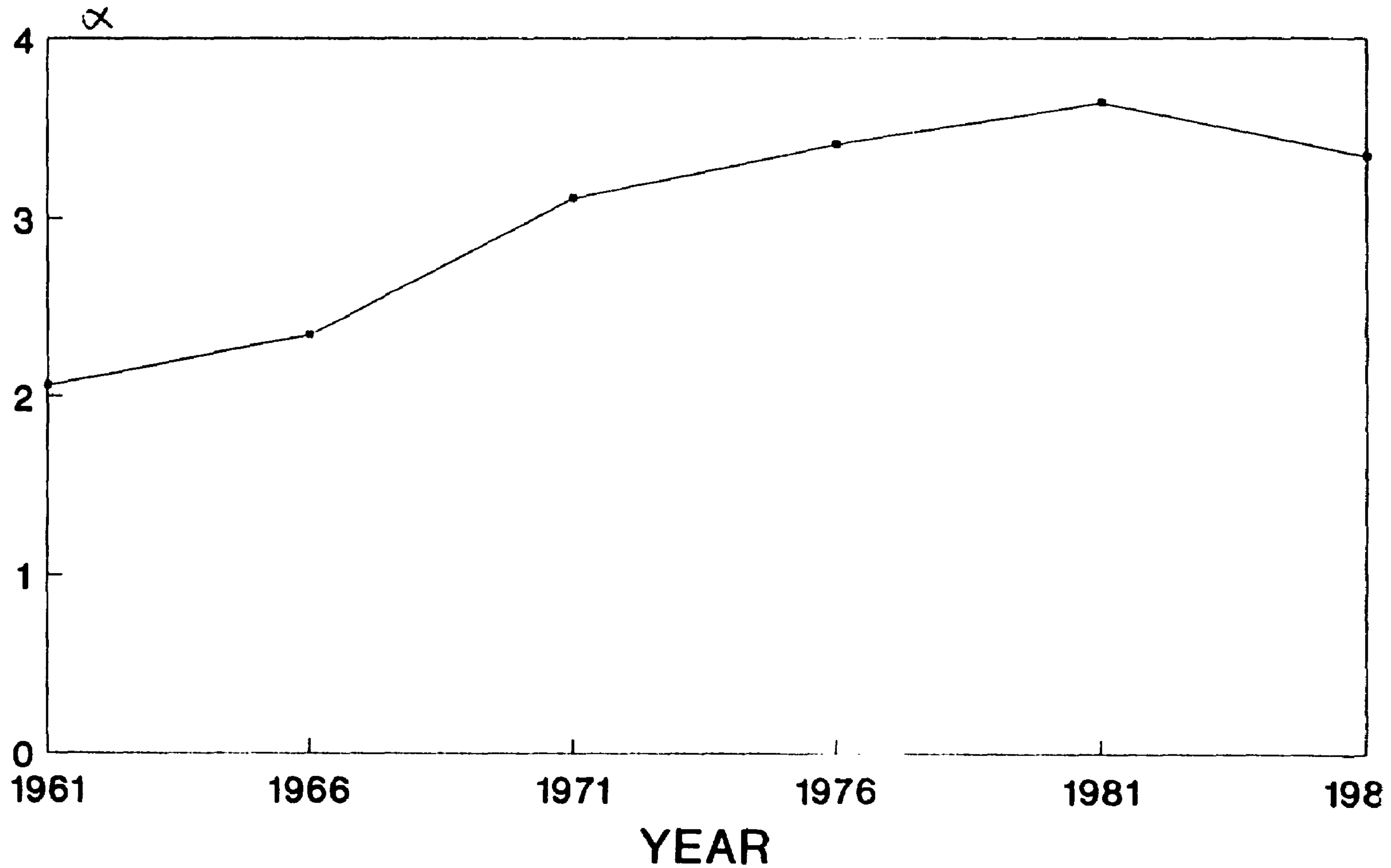


FIG. 5.2 : α VERSUS YEAR
MEDICAL SCIENCES

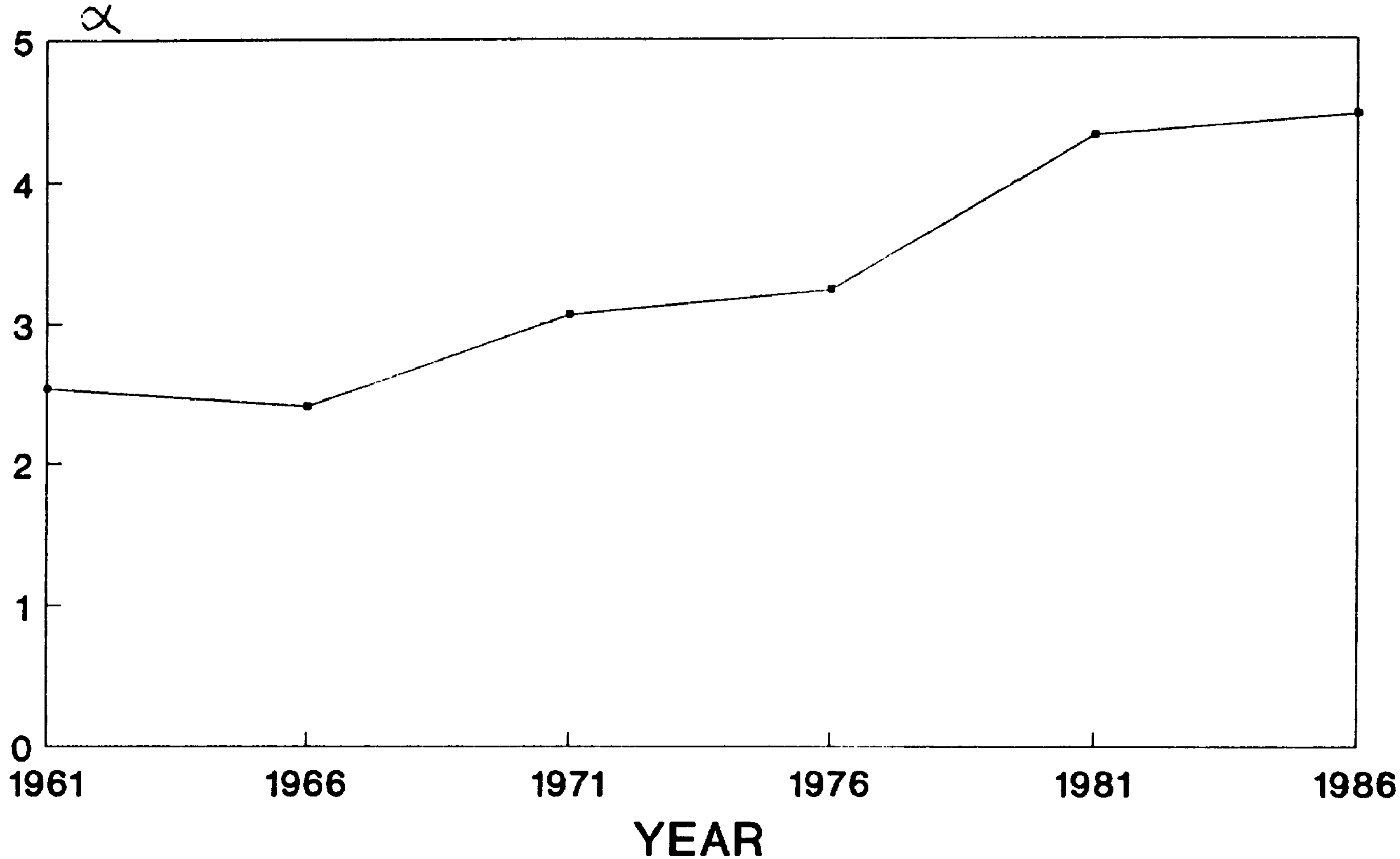


FIG. 5.3 : α VERSUS YEAR
PHYSICAL SCIENCES

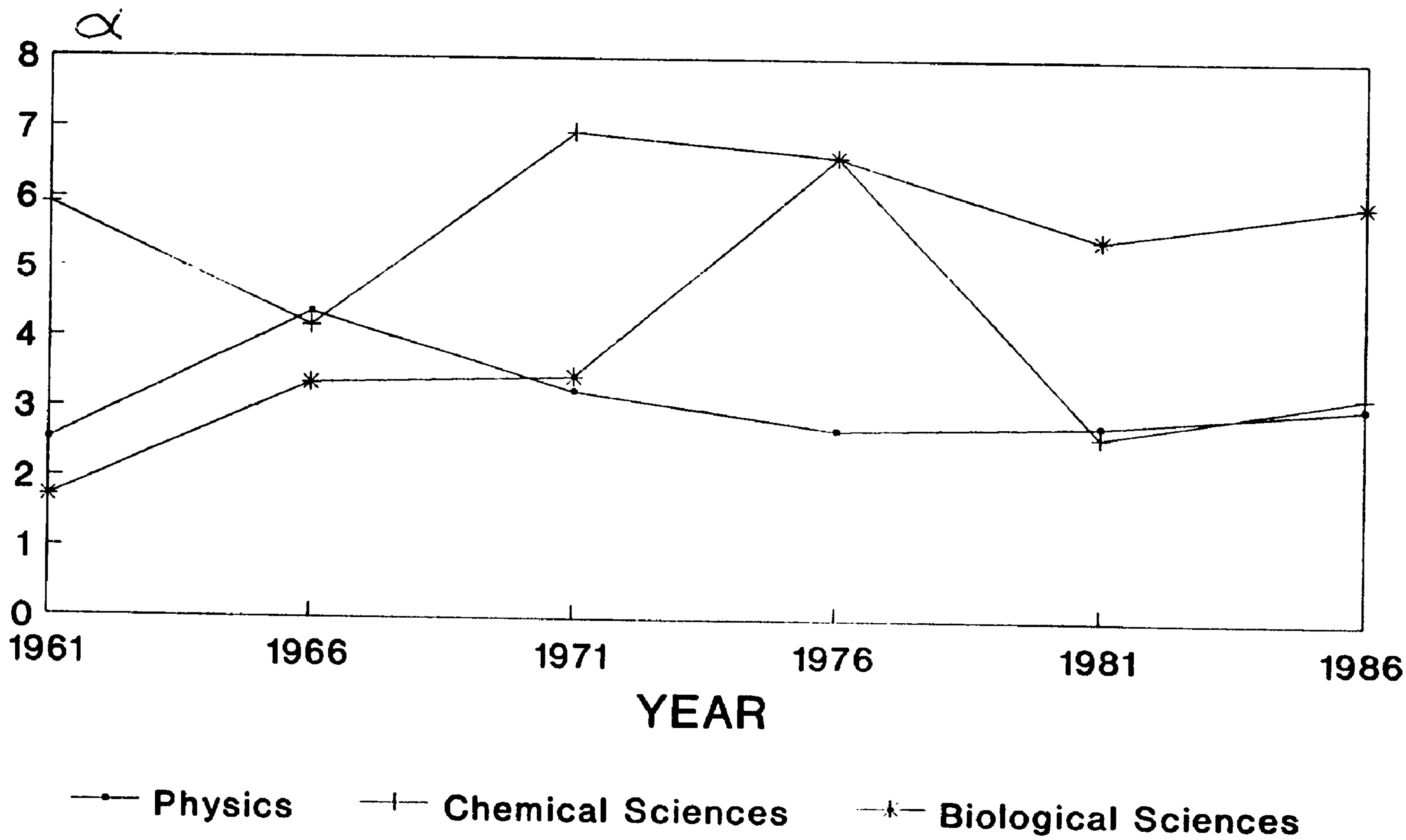
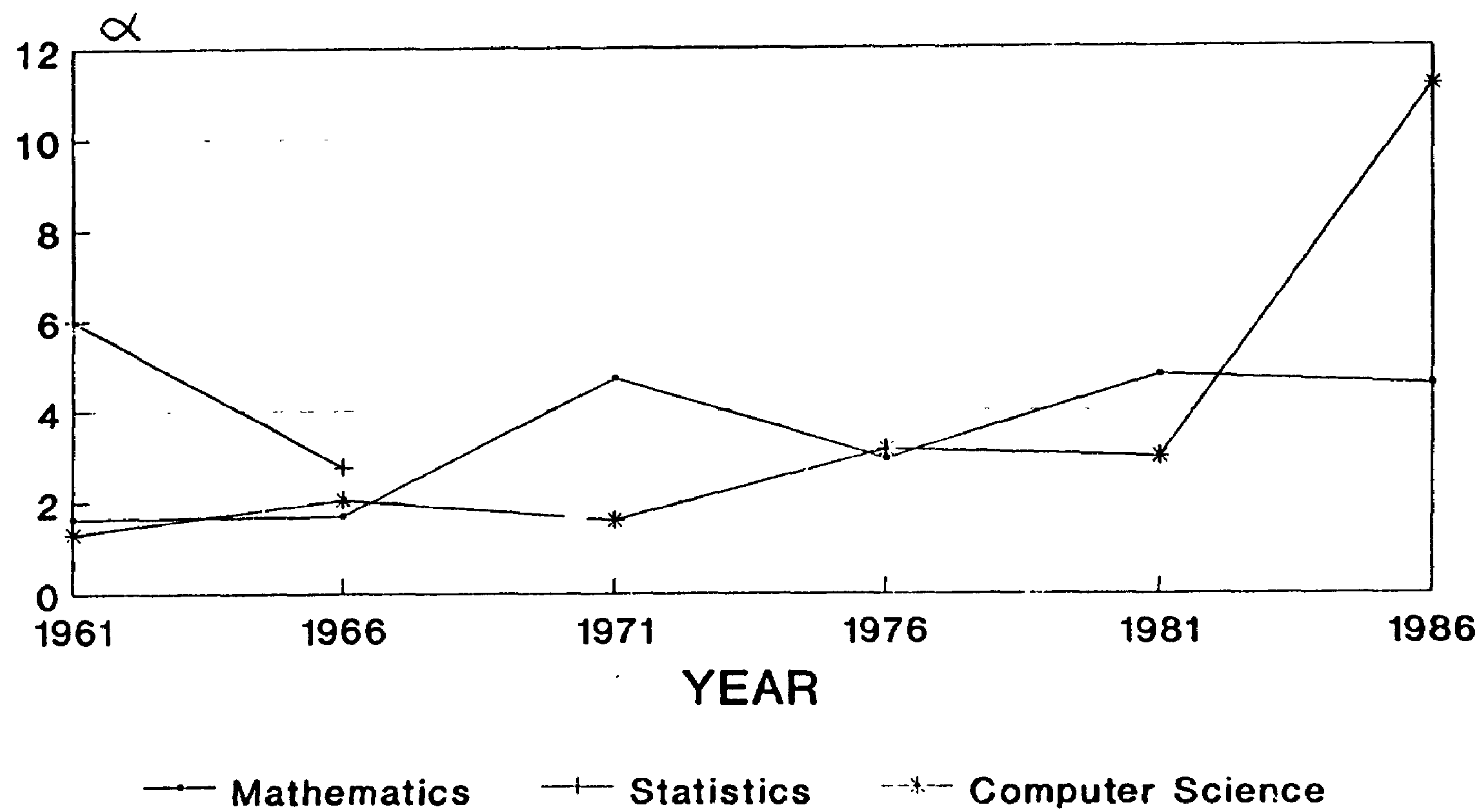


FIG. 5.4 : α VERSUS YEAR
MATHEMATICAL SCIENCES



No valid estimates for Statistics in
1971, 1976, 1981 and 1986

FIG. 5.5 : α VERSUS YEAR
SOCIAL SCIENCES

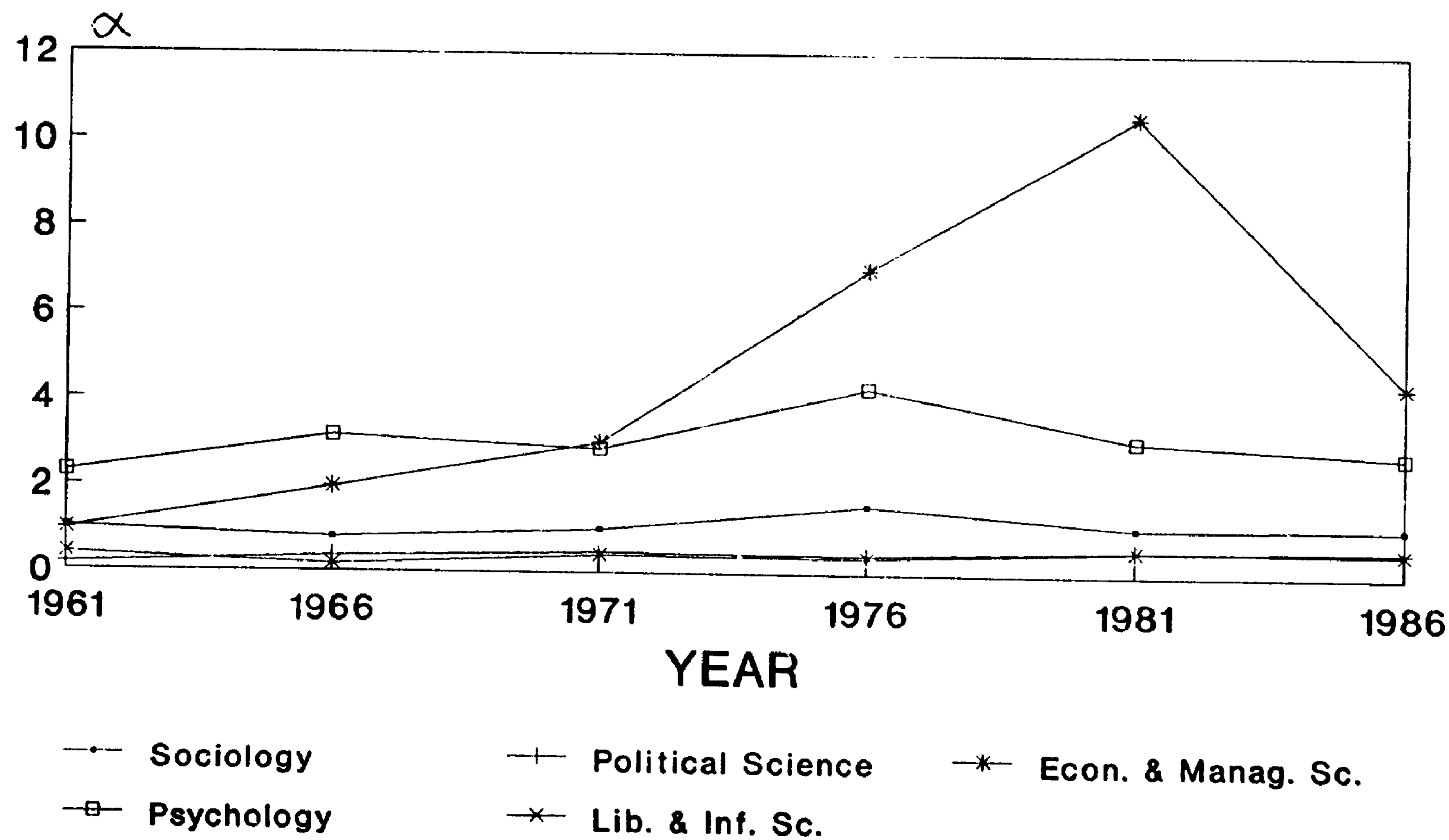
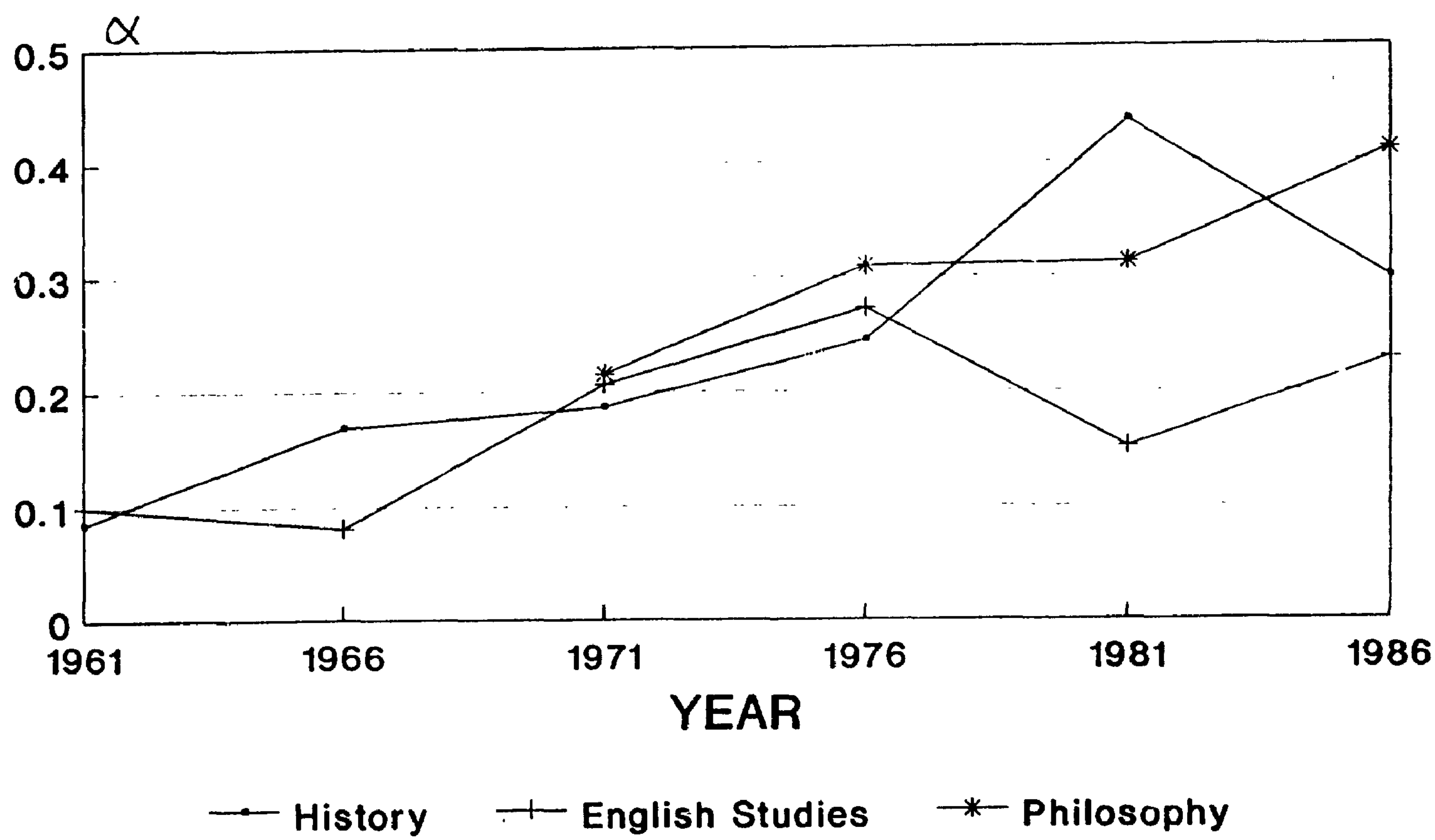


FIG. 5.6 : α VERSUS YEAR

HUMANITIES



No data for Philosophy in 1961 and 1966

TABLE OF CONTENTS

FIG. 5.7 : Θ VERSUS YEAR
ENGINEERING SCIENCES

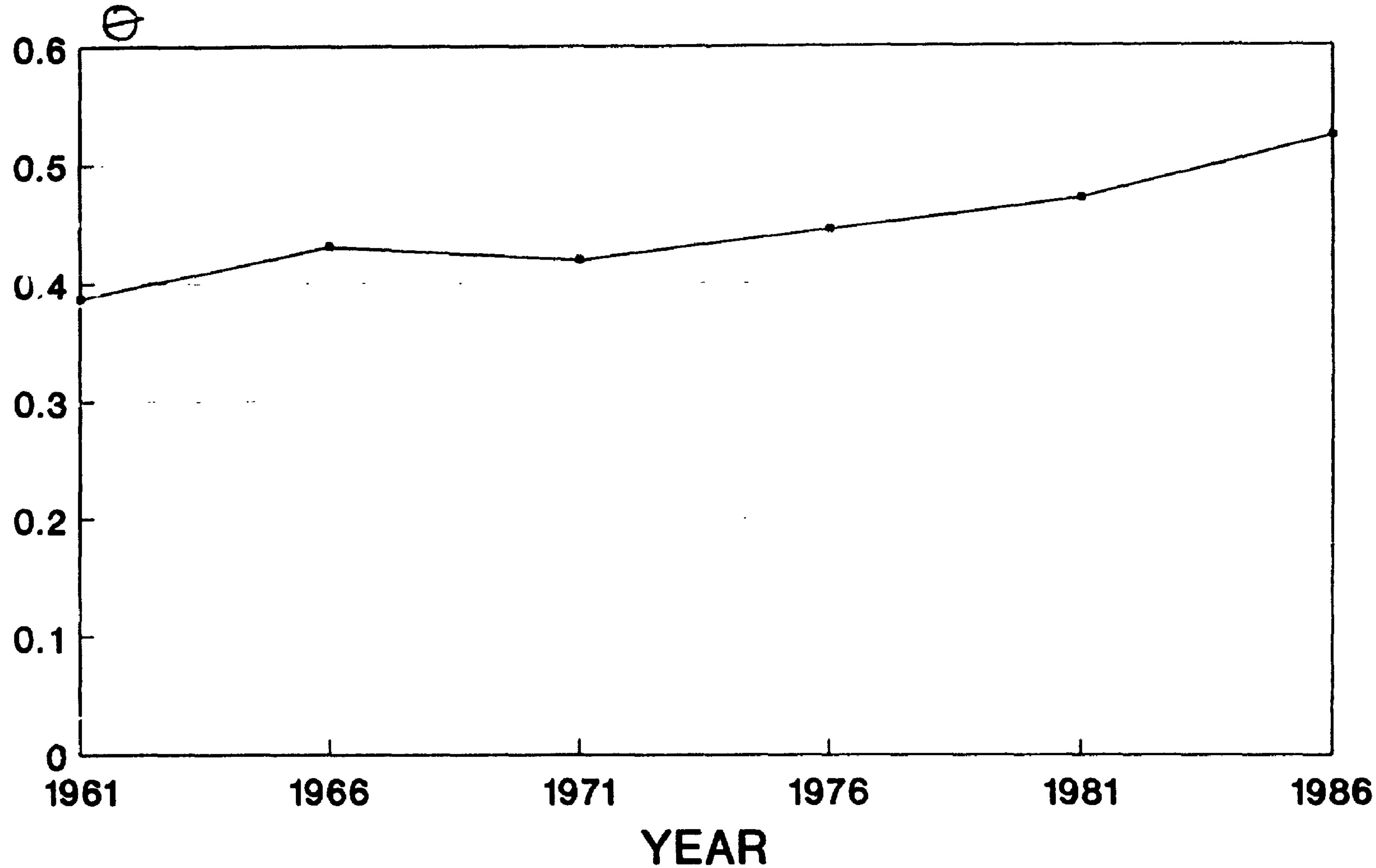


FIG. 5.8 : Θ VERSUS YEAR
MEDICAL SCIENCES

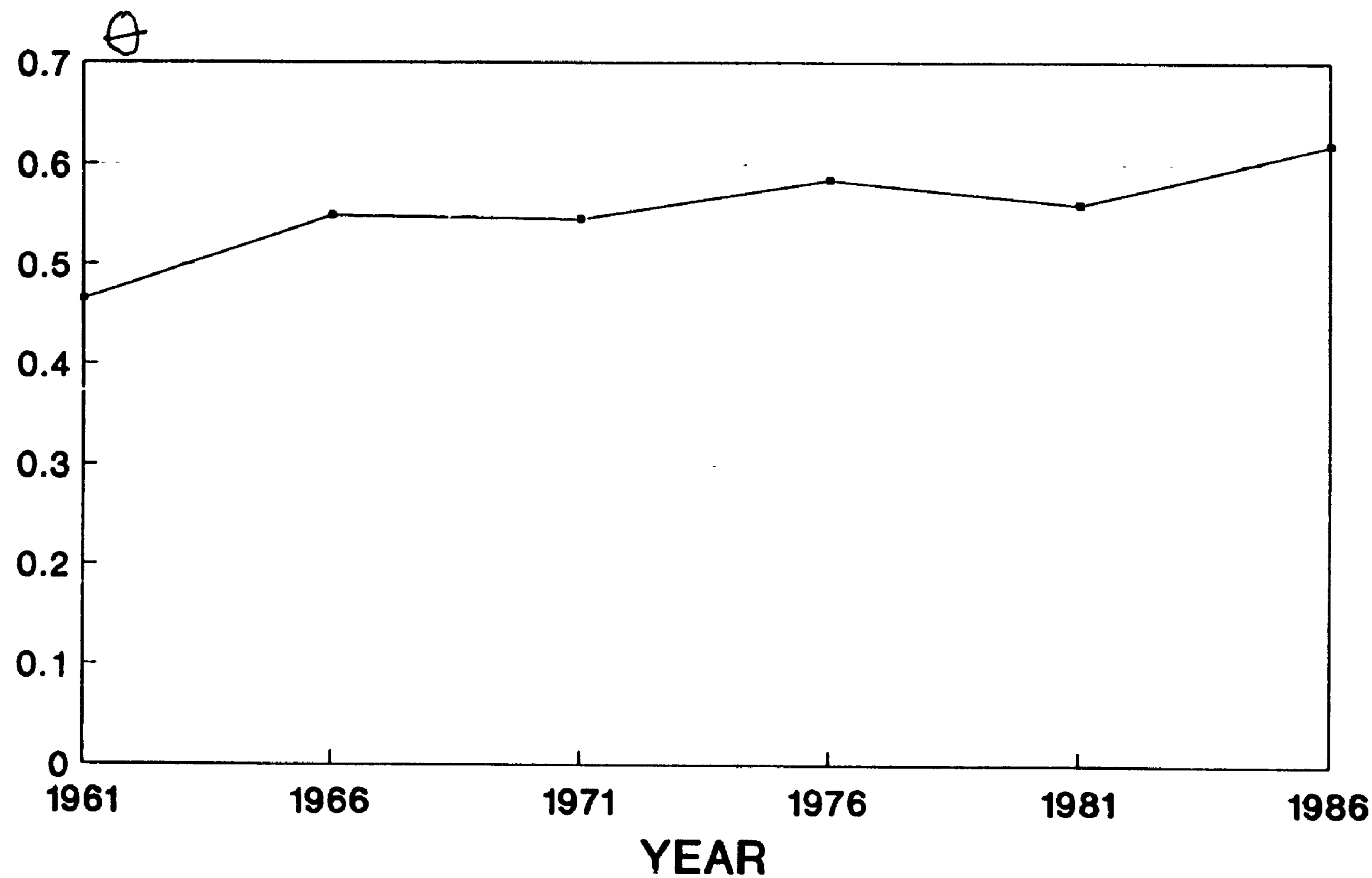


FIG. 5.9 : Θ VERSUS YEAR
PHYSICAL SCIENCES

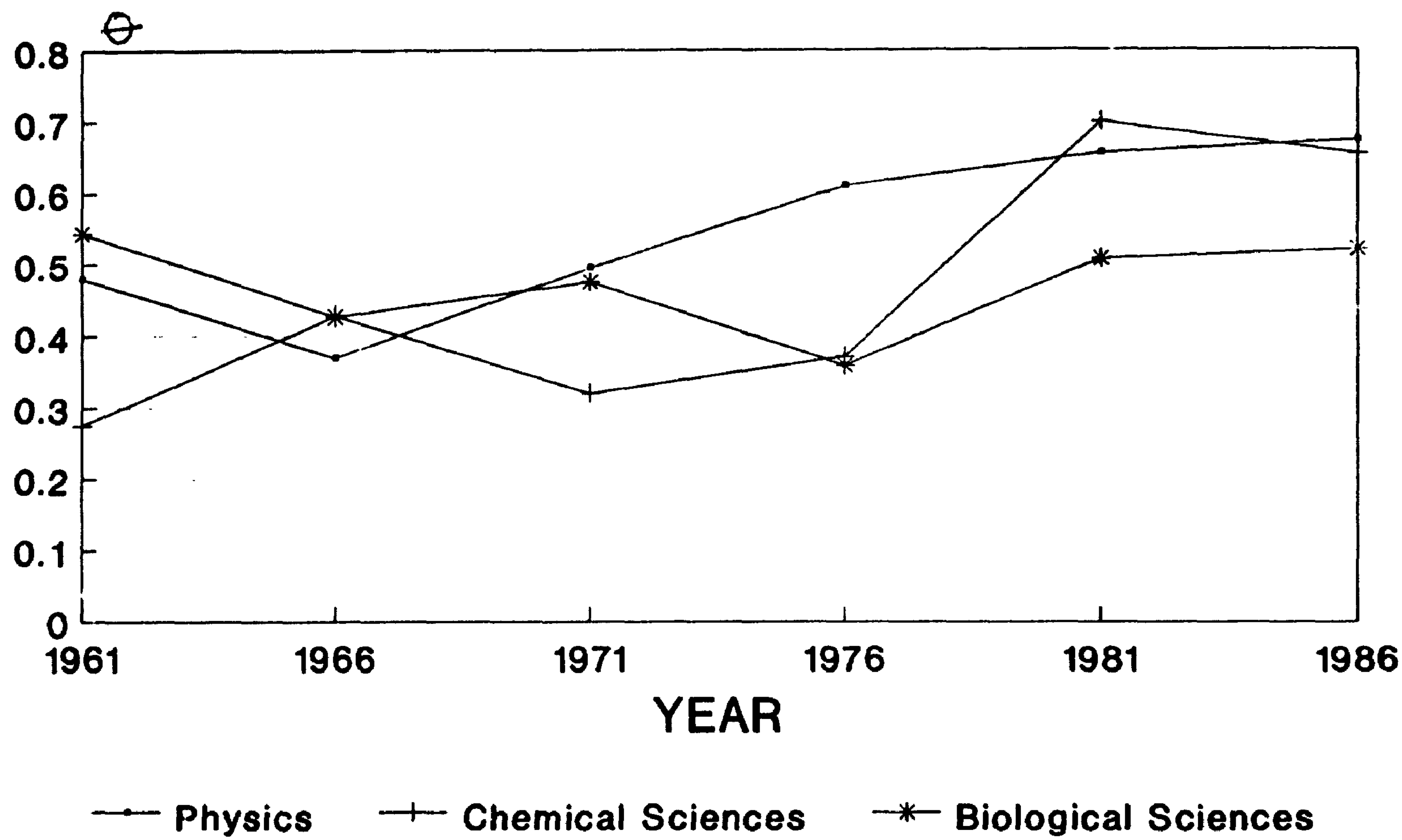
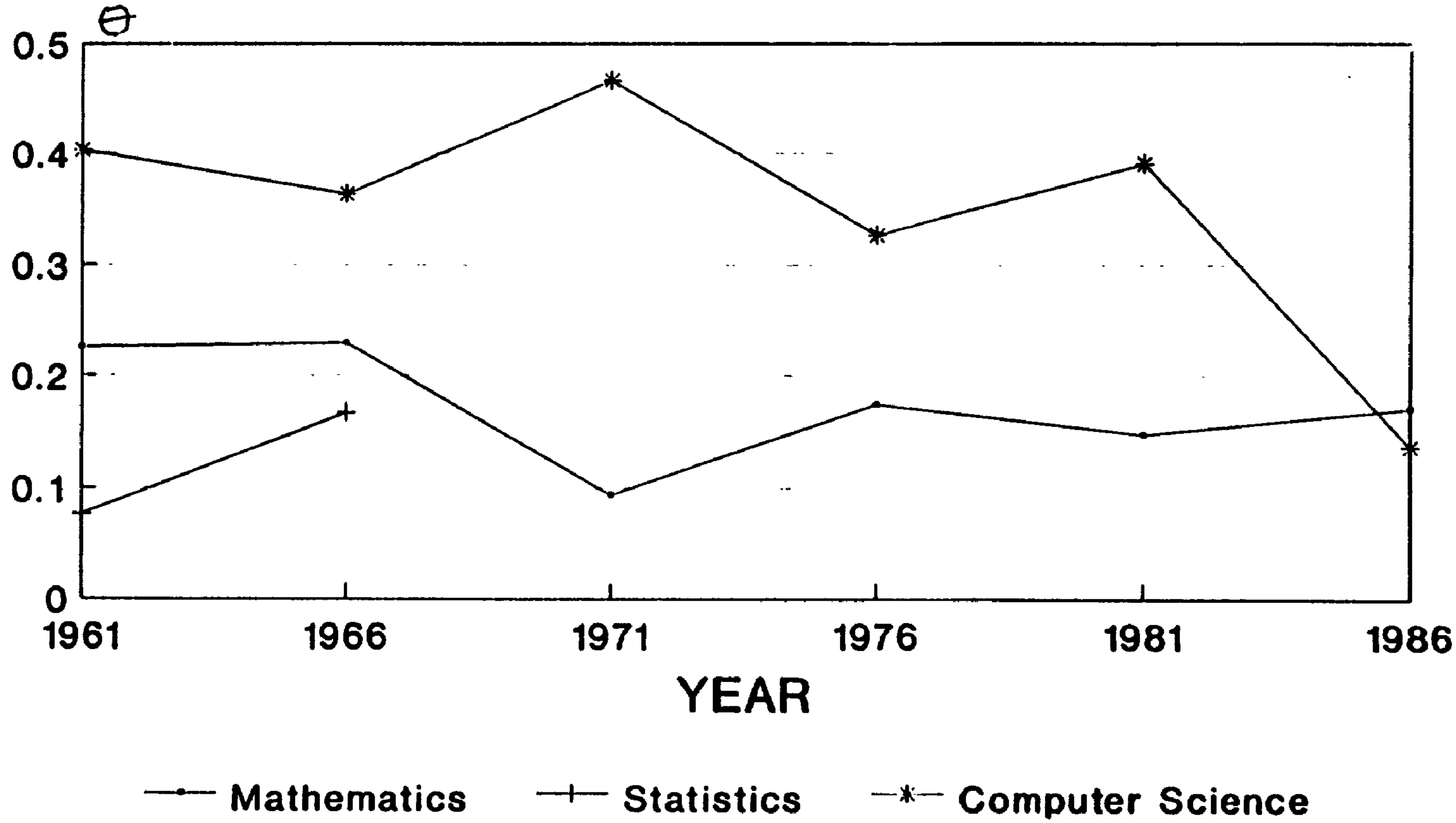


FIG. 5.10 : Θ VERSUS YEAR
MATHEMATICAL SCIENCES



No valid estimates for Statistics in
1971, 1976, 1981 and 1986

FIG. 5.11 : Θ VERSUS YEAR
SOCIAL SCIENCES

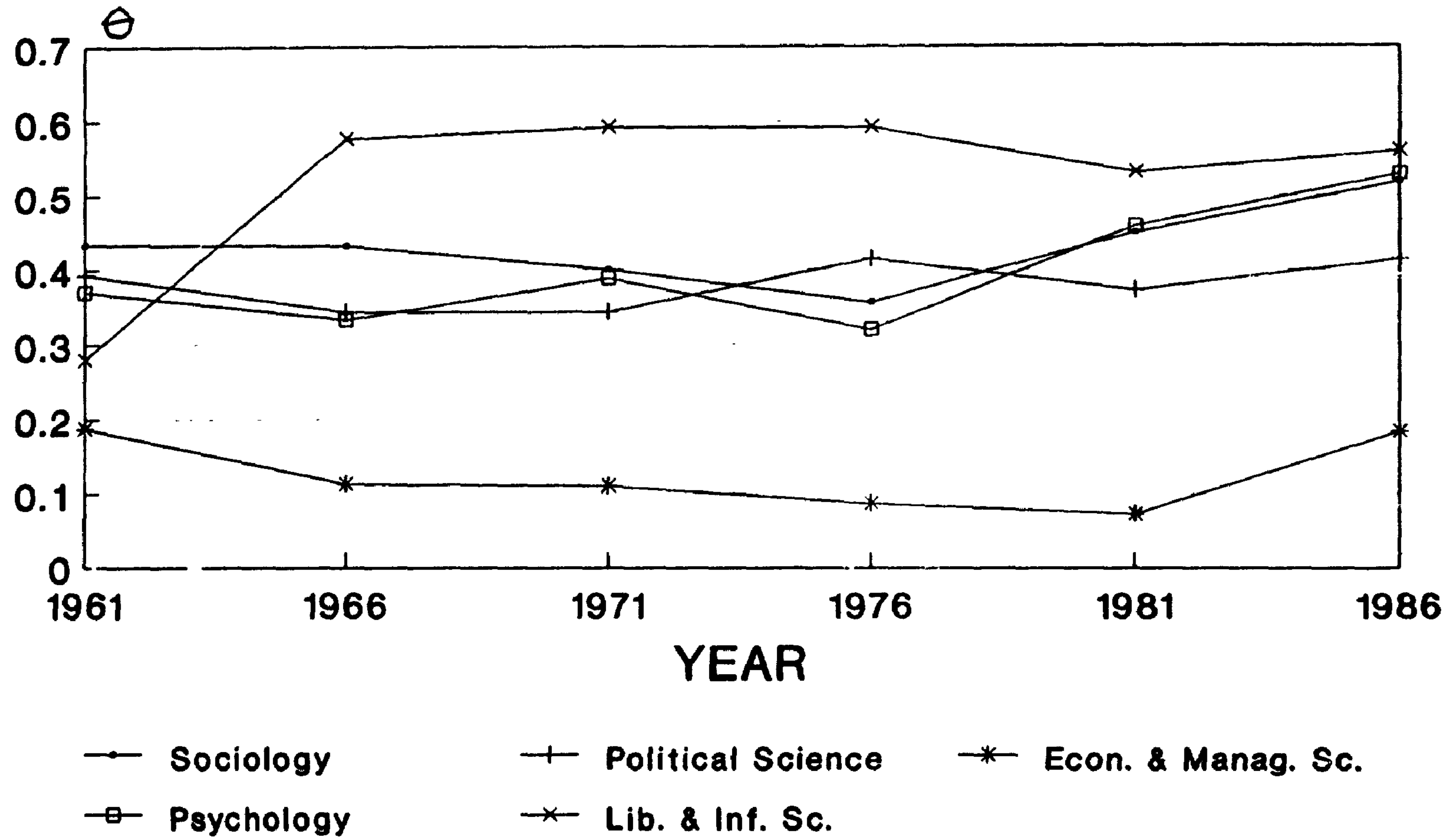
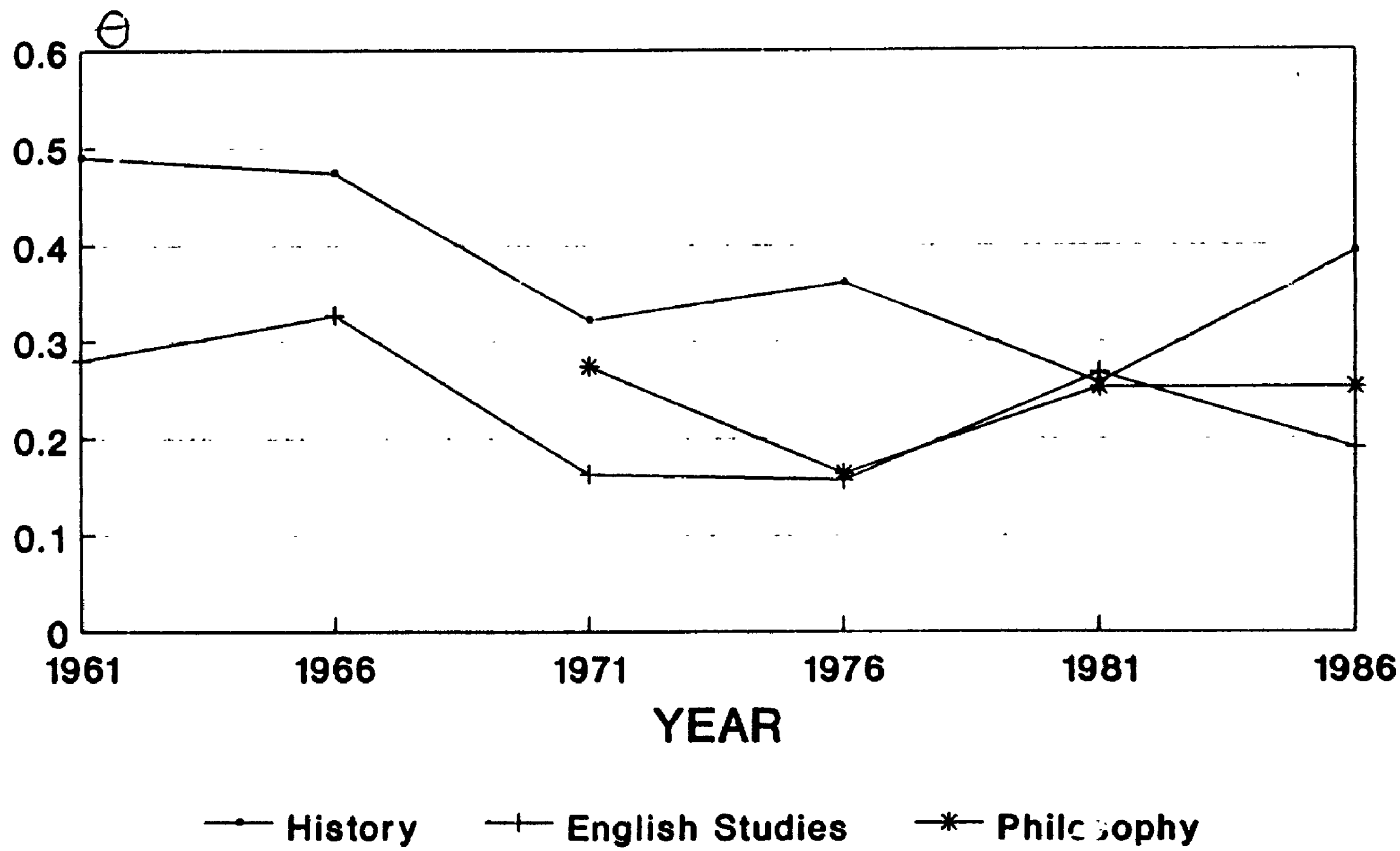


FIG. 5.12 : Θ VERSUS YEAR
HUMANITIES



No data for Philosophy in 1961 and 1966

5.5 Practical Applications

In addition to the advantage of being able to summarize the entire frequency distribution by a few parameters of the model, a model for the distribution of authorships can be used to estimate the number of entries in an author index. Abstracting services often add extra entries in the case of multiple authorships. For large abstracting journals and indexes, it would be time consuming to determine ahead of publication how many author entries would be made for N, a fixed number, papers. So, in order to estimate the number of entries, an abstractor may randomly sample n out of the N papers and estimate the parameters of the model from the sample. The values of the parameters could then be used to estimate for the N papers, the number of papers with one author, two authors, three authors, and so on. The abstractor then should be able to estimate the total number of entries, and, hence, have a rough estimate of the size of the author index. It is noted, though, that a rough estimate of the size of the author index could also be obtained from the sample estimate of the mean number of authors per paper.

The method described above can also be used to determine, subject to space constraints, the maximum number (m) of authors per paper to be included in an author index. Only the first author or the first m authors would be included for papers with more than m authors.

Example :

The Biological Abstracts indexed 120,102 papers in 1966. Suppose the manager of the abstracting service had space in the author index for only 200,000 entries and wanted to determine m, the maximum number of entries per paper. Assume he knew that the model for the distribution of authorships is the shifted inverse Gaussian-Poisson and that the estimates of the parameters from the sample taken are the same as the ones in Table 5.8. Then the expected distribution of authorships for the 120,102 papers is given below in Table 5.9

If the manager's policy was to include only the first author of a paper with more than m authors, then m is 3 and the estimated total number of entries is 182,305 + (1 X 11,084) = 193,389. However, if his policy was to include the first m authors of a paper with more than m authors, then m is 2 and the estimated total number of entries is 129,451 + (2 X 28,702) = 186,855 .

Table 5.9 : Expected distribution of authorships for the 120102 papers indexed in 1966 Biological Abstracts

No. of Authors	No. of Papers	No. of Author Entries	Cumulative No. of Author Entries
1	53349	53349	53349
2	38051	76102	129451
3	17618	52854	182305
4	6976	27904	210209
5	2603	13015	223224
6	953	5718	228942
7	348	2436	231378
8	128	1024	232402
9	47	423	232825
10	18	180	233005
11	8	88	233093
12	2	24	233117
13	1	13	233130
Total	120102	233130	

CHAPTER 6

RELATIONSHIPS BETWEEN THE MEASURES AND THE PARAMETERS OF THE MODELS

It is sometimes convenient if a relationship can be established between a measure of strength or inequality and a theoretical distribution which fits the observed distribution of a social phenomenon (Allison, 1980). In most cases, then, the measure can be estimated from the parameters of the distribution.

The collaborative coefficient along with the other three measures of collaboration is given below for three probability models that fitted at least half of the data sets. The derivation of E(1/X) for each of the models is given in Appendix G.

(i) Shifted Generalized Poisson

$$P(X = j) = \frac{\lambda_1[\lambda_1 + (j-1)\lambda_2]^{j-2} e^{-[\lambda_1 + (j-1)\lambda_2]}}{(j-1)!} \quad ; j = 1,2,..$$

$$P(X=1) = e^{-\lambda_1}$$

$$E(X) = \sum_j j\ P(x=j)$$

$$= 1 + \frac{\lambda_1}{1 - \lambda_2}$$

$$E[1/X] = \sum_j (1/j)\ P(x=j)$$

$$= \frac{\lambda_1 e^{\lambda_1-\lambda_2} - \lambda_1 - \lambda_1\lambda_2 e^{\lambda_1-\lambda_2} + \lambda_2{}^2 e^{\lambda_1-\lambda_2}}{(\lambda_1-\lambda_2)^2 \quad e^{\lambda_1-\lambda_2}}$$

Hence, CC = 1 - E[1/X]

$$= \frac{\lambda_1 [1 - \lambda_2 e^{\lambda_1-\lambda_2} + \lambda_1 e^{\lambda-\lambda_2} - e^{\lambda_1-\lambda_2}]}{(\lambda_1-\lambda_2)^2 \quad e^{\lambda_1-\lambda_2}}$$

CI = E[X]

$$= 1 + \frac{\lambda_1}{1-\lambda_2}$$

DC = 1 - P(X=1)

$$= 1 - e^{-\lambda_1}$$

Note that :

CC → 0, CI → 1 and DC → 0 as $\lambda_1 \to 0$; and

CC → 1, CI → ∞ and DC → 1 as $\lambda_1 \to \infty$

(ii) Shifted Inverse Gaussian-Poisson

$$P(X = j) = (2\alpha/\pi)^{1/2} e^{\alpha(1-\theta)^{1/2}} \frac{(\frac{1}{2}\alpha\theta)^{j-1}}{(j-1)!} K_{j-3/2}(\alpha) \quad ; \; j = 1,2,...$$

$$P(X=1) = e^{\alpha[(1-\theta)^{1/2} - 1]}$$

$$E[X] = \sum_j j \; P(x=j)$$

$$= \frac{\alpha\theta}{2(1-\theta)^{1/2}} + 1$$

$$E[1/X] = \Sigma \; (1/j) \; P(x=j)$$

$$= \frac{2[\alpha(1-\theta)^{1/2}+1]}{\alpha^2\theta} \; [1 - ((\alpha+1)/\alpha)(\pi/2\alpha)^{1/2} \; e^{-\alpha}]$$

Hence, CC = 1 - E[1/X]

$$= 1 - \frac{2[(1-\theta)^{\frac{1}{2}}+1]}{\alpha^2 \theta} \quad [1 - ((\alpha+1)/\alpha)(\pi/2\alpha)^{\frac{1}{2}} e^{-\alpha}]$$

$$CI = E[X] = \frac{\alpha\theta}{2(1-\theta)^{\frac{1}{2}}} + 1$$

$$DC = 1 - P(X=1)$$

$$= 1 - e^{\alpha[(1-\theta)^{1/2} - 1]}$$

Note that :

$CC \to 1$, $CI \to \infty$ and $DC \to 1$ as $\theta \to 1$ and $\alpha \to \infty$; and

$CC \to 0$, $CI \to 1$ and $DC \to 0$ as $\theta \to 0$ and $\alpha \to \infty$

(iii) Shifted Negative Binomial

$$P(X = j) = \binom{v+j-2}{j-1} p^v (1-p)^{j-1} \quad ; \; j = 1,2, \ldots$$

$$P(X=1) = p^v$$

$$E[X] = \sum_j j \binom{v+j-2}{j-1} p^v (1-p)^{j-1}$$

$$= \frac{v(1-p)}{p} + 1$$

$$E[1/X] = \sum_j (1/j) \binom{v+j-2}{j-1} p^v (1-p)^{j-1}$$

$$= \frac{p}{(1-p)(v-1)} (1 - p^{v-1})$$

Hence, $CC = 1 - E[1/X] = 1 - \dfrac{p}{(1-p)(v-1)} (1 - p^{v-1})$

$$CI = E[X] = \frac{v(1-p)}{p} + 1$$

$$DC = 1 - P(X=1) = 1 - p^v$$

Note that :

$CC \rightarrow 1$, $CI \rightarrow \infty$ and $DC \rightarrow 1$ as $p \rightarrow 0$; and

$CC \rightarrow 0$, $CI \rightarrow 1$ and $DC \rightarrow 0$ as $p \rightarrow 1$

CHAPTER 7

CONCLUSIONS

The collaborative coefficient (CC) has been derived as an alternative to either the mean number of authors or the proportion of multiple-authorships as a measure of the degree of research collaboration.

Using CC as a measure, the humanists are found to be the least collaborative with an average value of 0.0172 over the three decades while the engineering, medical and physical scientists are the most collaborative with average values of 0.3154, 0.4081 and 0.4169 respectively. It is also found that collaboration has been increasing steadily from the '60s to the '80s for all fields. The average CC value of 0.1484 for all fields in the 60's increased to 0.2495 in the 80's.

The shifted Waring distribution was derived as the probability model for the distribution of authorships but empirical data showed that the shifted inverse Gaussian-Poisson is a better model. This can be derived as an appropriate model for the distribution of authorships (see Appendix F). It is noted, though, that the model didn't fit very well the data sets from the most collaborative fields, i.e., engineering, physical and medical sciences. However, until a better model is developed, it is suggested that the shifted inverse Gaussian-Poisson be assumed as the model for the distribution of authorships. This model could be used in estimating the number of entries in an author index and in determining the maximum number authors per paper to be included in an author index.

APPENDIX A : DATA SETS

Table A.1 : Distribution of Authorships for Engineering Sciences

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1974	2473	3100	2895	1837	3050
2	785	1299	2153	2218	1705	2500
3	241	481	899	1166	962	1682
4	61	123	318	403	371	792
5	16	37	92	180	144	274
6	8	15	28	54	62	130
7	2	7	12	27	28	58
8	1	1	8	11	9	21
9	-	0	4	5	5	10
10	-	1	5	2	3	7
11	-	1	1	5	2	6
12	-	1	2	0	3	3
13	-	2	0	0	1	2
14	-	1	2	0	1	0
15	-	-	-	1	1	3
16	-	-	-	-	3	6
Total	3088	4442	6624	6967	5137	8544
Mean	1.5094	1.6684	1.8569	2.0205	2.1803	2.2726
Variance	0.6718	0.9846	1.2206	1.4602	1.8203	2.0679

Table A.2 : Distribution of Authorships for Medical Sciences

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1365	1366	1526	1837	1351	1204
2	729	780	1184	1432	1244	1252
3	362	441	734	1073	1054	1123
4	131	203	362	671	727	877
5	38	73	140	252	461	652
6	13	39	58	103	177	330
7	8	8	19	52	98	166
8	2	4	24	29	39	86
9	1	2	6	23	19	54
10	-	1	5	11	4	24
11	-	1	3	6	5	12
12	-	-	2	3	3	10
13	-	-	-	2	1	6
14	-	-	-	1	0	2
15	-	-	-	-	1	1
16	-	-	-	-	0	1
17	-	-	-	-	0	-
18	-	-	-	-	0	-
19	-	-	-	-	0	-
20	-	-	-	-	0	-
21	-	-	-	-	1	-
Total	2649	2918	4063	5495	5185	5800
Mean	1.8052	1.9832	2.2343	2.4635	2.8206	3.2397
Variance	1.1410	1.5379	1.9864	2.4893	2.9319	3.9401

Table A.3 : Distribution of Authorships for Physics

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	537	1217	1980	1551	1352	1066
2	370	1062	1960	1653	1446	1328
3	140	548	910	868	1006	944
4	38	158	328	392	451	535
5	21	55	128	130	189	249
6	8	19	51	54	95	128
7	3	13	30	39	38	62
8	2	5	19	8	20	32
9	0	4	5	11	16	22
10	0	0	6	8	10	11
11	1	0	4	5	7	8
12	0	0	4	7	2	10
13	1	0	3	3	5	6
14	-	1	3	3	5	4
15	-	0	0	3	2	2
16	-	0	0	3	1	2
17	-	0	0	1	1	1
18	-	0	0	1	0	1
19	-	0	0	0	1	0
20	-	0	1	1	3	1
21	-	0	-	1	0	0
22	-	0	-	0	1	0
23	-	0	-	1	1	0
24	-	0	-	2	0	0
25	-	0	-	1	0	0
30	-	0	-	0	0	1
33	-	0	-	0	0	2
34	-	0	-	0	1	0
36	-	0	-	0	1	0
38	-	0	-	0	1	0
41	-	0	-	0	1	0
43	-	0	-	0	1	1
47	-	0	-	0	0	1
48	-	0	-	1	0	0
51	-	1	-	-	0	0
56	-	-	-	-	0	2
62	-	-	-	-	1	0
69	-	-	-	-	0	1
91	-	-	-	-	1	-
97	-	-	-	-	1	-
Total	1121	3083	5432	4747	4660	4420
Mean	1.8403	2.0234	2.1259	2.3282	2.6000	2.8351
Variance	1.3093	2.0897	1.8394	3.3790	8.7951	7.0285

Table A.4 : Distribution of Authorships for Chemical Sciences

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1691	2475	1263	1556	1894	1818
2	1559	2527	1611	1988	2068	2312
3	682	1397	995	1408	1633	1932
4	211	508	448	755	852	1061
5	62	196	142	315	391	475
6	18	63	66	120	188	223
7	11	30	19	63	71	93
8	7	18	8	26	41	46
9	1	3	4	8	20	25
10	1	4	5	4	6	17
11	1	7	3	3	8	14
12	0	4	1	2	7	15
13	0	3	1	1	6	10
14	0	2	-	-	7	9
15	1	2	-	-	4	6
16	-	2	-	-	2	4
17	-	1	-	-	1	2
18	-	-	-	-	1	3
19	-	-	-	-	2	2
20	-	-	-	-	3	2
21	-	-	-	-	1	0
22	-	-	-	-	2	0
23	-	-	-	-	1	0
24	-	-	-	-	2	0
25	-	-	-	-	0	0
26	-	-	-	-	0	1
27	-	-	-	-	0	1
28	-	-	-	-	0	0
29	-	-	-	-	0	0
30	-	-	-	-	0	0
35	-	-	-	-	1	0
55	-	-	-	-	0	1
61	-	-	-	-	1	0
62	-	-	-	-	1	0
79	-	-	-	-	1	0
Total	4245	7241	4566	6249	7215	8072
Mean	1.9541	2.1801	2.3454	2.5447	2.6931	2.8031
Variance	1.1649	1.7430	1.6610	2.0179	5.1928	3.7562

Table A.5 : Distribution of Authorships for Biological Sciences

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1672	2135	2223	1489	1311	1367
2	796	1459	1822	1920	2050	2643
3	308	697	1009	1212	1598	2303
4	98	316	433	646	1038	1559
5	35	99	175	273	515	900
6	8	34	68	111	298	483
7	6	8	21	44	127	247
8	5	5	9	24	58	119
9	3	0	3	8	33	60
10	2	0	3	3	16	35
11	2	0	3	0	23	25
12	1	1	1	0	6	25
13	-	0	-	1	3	8
14	-	0	-	-	2	3
15	-	1	-	-	1	7
16	-	-	-	-	2	1
17	-	-	-	-	-	1
18	-	-	-	-	-	5
24	-	-	1	-	-	-
27	-	-	1	-	-	-
Total	2936	4755	5772	5731	7081	9791
Mean	1.6914	1.9411	2.1277	2.4769	2.9613	3.2393
Variance	1.1593	1.2965	1.7628	1.8994	3.0366	3.5898

Table A.6 : Distribution of Authorships for Mathematics

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	10355	12215	14143	23267	23403	25535
2	1926	2418	3198	6148	8628	10399
3	279	325	390	945	1660	2313
4	36	55	40	133	248	385
5	7	6	6	21	41	70
6	1	7	3	7	16	16
7	0	1	0	5	5	7
8	0	1	1	1	1	2
9	0	-	-	2	2	1
10	0	-	-	1	1	3
11	0	-	-	-	1	2
12	0	-	-	-	0	0
13	1	-	-	-	0	0
14	-	-	-	-	0	1
15	-	-	-	-	0	0
16	-	-	-	-	0	1
17	-	-	-	-	1	-
Total	12605	15028	17781	30530	34007	38735
Mean	1.2092	1.2199	1.2331	1.2823	1.3830	1.4306
Variance	0.2456	0.2557	0.2459	0.3158	0.4223	0.4829

Table A.7 : Distribution of Authorships for Statistics

Number of Authors	Number of Papers					
	1961	1966	1971	1976	1981	1986
1	527	635	904	1248	1722	2171
2	120	151	321	524	832	1275
3	16	18	49	85	141	261
4	2	4	3	6	17	36
5	-	1	-	-	2	4
6	-	-	-	-	-	1
7	-	-	-	-	-	3
Total	665	809	1277	1863	2714	3751
Mean	1.2376	1.2509	1.3352	1.3822	1.4322	1.5183
Variance	0.2477	0.2773	0.3139	0.3469	0.3959	0.4887

Table A.8 : Distribution of Authorships for Computer Science

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1252	1436	1382	1235	1433	1263
2	314	544	523	693	930	1123
3	79	129	160	201	318	403
4	25	53	41	55	90	85
5	3	8	14	14	38	25
6	1	4	4	7	13	14
7	-	-	5	3	3	3
8	-	-	2	0	2	2
9	-	-	-	2	1	2
10	-	-	-	-	0	-
11	-	-	-	-	1	-
12	-	-	-	-	0	-
13	-	-	-	-	0	-
14	-	-	-	-	0	-
15	-	-	-	-	0	-
16	-	-	-	-	1	-
Total	1674	2174	2131	2210	2830	2920
Mean	1.3369	1.4660	1.5096	1.6267	1.7484	1.8226
Variance	0.4411	0.5950	0.7420	0.7963	1.0445	0.8935

Table A.9 : Distribution of Authorships for Sociology

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1572	3802	4910	5217	6669	6856
2	348	649	972	1443	1614	1962
3	70	134	212	353	431	555
4	16	27	41	83	100	161
5	11	6	12	15	20	57
6	1	2	3	3	12	17
7	-	2	2	1	2	9
8	-	0	-	2	0	4
9	-	1	-	-	0	1
10	-	-	-	-	2	1
11	-	-	-	-	1	1
12	-	-	-	-	-	1
Total	2018	4623	6152	7117	8851	9625
Mean	1.2899	1.2276	1.2591	1.3503	1.3340	1.4143
Variance	0.3983	0.3182	0.3438	0.4466	0.4751	0.6444

Table A.10 : Distribution of Authorships for Political Science

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1421	1388	2043	4490	5020	5191
2	50	86	170	407	560	648
3	4	12	19	61	72	106
4	1	2	7	17	27	22
5	1	-	-	2	1	4
6	-	-	-	1	1	3
7	-	-	-	-	-	0
8	-	-	-	-	-	0
9	-	-	-	-	-	0
10	-	-	-	-	-	0
11	-	-	-	-	0	1
Total	1477	1488	2239	4978	5681	5975
Mean	1.0440	1.0780	1.1023	1.1191	1.1398	1.1618
Variance	0.0597	0.0961	0.1276	0.1588	0.1798	0.2264

Table A.11 : Distribution of Authorships for Economic Science and Managerial Sciences

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	1546	1810	1415	1565	1143	235
2	134	203	220	473	436	110
3	17	16	35	90	102	16
4	-	2	-	8	8	3
5	-	-	-	2	4	2
6	-	-	-	-	-	1
Total	1697	2031	1677	2138	1693	367
Mean	1.0990	1.1187	1.1771	1.3204	1.4017	1.4469
Variance	0.1093	0.1263	0.1876	0.3358	0.4178	0.5047

Table A.12 : Distribution of Authorships for Psychology

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	4097	6586	4754	2590	1965	2185
2	1701	3479	2532	1831	1476	1474
3	553	1231	1035	747	635	806
4	155	342	336	239	275	374
5	39	87	112	78	103	156
6	11	29	23	31	33	66
7	2	11	6	9	16	20
8	1	7	3	2	10	10
9	0	1	2	1	3	4
10	0	-	0	-	1	5
11	1	-	1	-	1	1
12	-	-	-	-	1	-
Total	6560	11773	8804	5528	4519	5101
Mean	1.5354	1.6441	1.7106	1.8294	1.9666	2.0663
Variance	0.6934	0.8084	0.9325	1.0270	1.3949	1.6435

Table A.13 : Distribution of Authorships for Library and Information Science

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	783	1021	1968	2771	3697	4971
2	43	48	232	312	559	786
3	6	10	54	65	123	170
4	-	3	15	23	33	36
5	-	1	8	6	8	17
6	-	-	1	5	5	10
7	-	-	0	1	4	3
8	-	-	1	-	-	0
9	-	-	-	-	-	2
10	-	-	-	-	-	3
Total	832	1083	2279	3183	4429	5998
Mean	1.0660	1.0748	1.1880	1.1778	1.2224	1.2356
Variance	0.0763	0.1155	0.3092	0.2940	0.3446	0.4099

Table A.14 : Distribution of Authorships for History

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	2455	3151	5943	8319	9255	17019
2	53	130	182	370	518	1021
3	4	12	13	37	49	104
4	3	5	1	9	8	34
5	1	2	2	3	-	4
6	-	1	1	-	-	2
7	-	-	-	-	-	0
8	-	-	-	-	-	1
9	-	-	-	-	-	0
10	-	-	-	-	-	1
Total	2516	3301	6142	8738	9830	18186
Mean	1.0294	1.0551	1.0365	1.0553	1.0651	1.0755
Variance	0.0437	0.0818	0.0475	0.0710	0.0757	0.1036

Table A.15 : Distribution of Authorships for English Studies

Number of Authors	Number of Papers 1961	1966	1971	1976	1981	1986
1	2693	3453	3058	3144	2354	2553
2	38	46	52	68	49	55
3	2	4	1	2	2	4
4	1	1	1	1	0	-
5	-	-	-	-	1	-
Total	2734	3504	3112	3215	2406	2612
Mean	1.0165	1.0163	1.0183	1.0233	1.0237	1.0241
Variance	0.0199	0.0200	0.0206	0.0259	0.0298	0.0266

Table A.16 : Distribution of Authorships for Philosophy

Number of Authors	Number of Papers					
	1967/68	1969	1971	1976	1981	1986
1	4086	2315	2968	3698	6145	7079
2	119	68	86	93	239	361
3	13	1	10	6	27	44
Total	4218	2384	3064	3797	6411	7484
Mean	1.0344	1.0294	1.0346	1.0277	1.0457	1.0600
Variance	0.0394	0.0294	0.0399	0.0301	0.0520	0.0682

APPENDIX B : MEASURES OF COLLABORATION

Table B.1 : Measures of Collaboration for Engineering Sciences

Year	CI	DC	CC
1961	1.5094	0.3608	0.2011
1966	1.6684	0.4433	0.2514
1971	1.8569	0.5320	0.3082
1976	2.0205	0.5845	0.3476
1981	2.1803	0.6424	0.3870
1986	2.2726	0.6430	0.3973

Table B.2 : Measures of Collaboration for Medical Sciences

Year	CI	DC	CC
1961	1.8052	0.4847	0.2849
1966	1.9832	0.5319	0.3225
1971	2.2343	0.6244	0.3851
1976	2.4635	0.6657	0.4246
1981	2.8206	0.7394	0.4889
1986	3.2397	0.7924	0.5423

Table B.3 : Measures of Collaboration for Physics

Year	CI	DC	CC
1961	1.8403	0.5210	0.3001
1966	2.0234	0.6053	0.3554
1971	2.2253	0.6355	0.3762
1976	2.3282	0.6733	0.4079
1981	2.6000	0.7099	0.4443
1986	2.8351	0.7588	0.4867

Table B.4 : Measures of Collaboration for Chemical Sciences

Year	CI	DC	CC
1961	1.9541	0.6016	0.3478
1966	2.1801	0.6582	0.3938
1971	2.3454	0.7234	0.4401
1976	2.5447	0.7510	0.4711
1981	2.6931	0.7375	0.4711
1986	2.8031	0.7748	0.4990

Table B.5 : Measures of Collaboraticn for Biological Sciences

Year	CI	DC	CC
1961	1.6914	0.4305	0.2480
1966	1.9411	0.5510	0.3263
1971	2.1277	0.6149	0.3711
1976	2.4769	0.7402	0.4594
1981	2.9613	0.8149	0.5319
1986	3.2393	0.8604	0.5738

Table B.6 : Measures of Collaboration for Mathematics

Year	CI	DC	CC
1961	1.2092	0.1785	0.0939
1966	1.2199	0.1872	0.0984
1971	1.2331	0.2046	0.1067
1976	1.2823	0.2379	0.1256
1981	1.3830	0.3118	0.1665
1986	1.4306	0.3408	0.1837

Table B.7 : Measures of Collaboration for Statistics

Year	CI	DC	CC
1961	1.2376	0.2075	0.1085
1966	1.2509	0.2151	0.1129
1971	1.3352	0.2921	0.1530
1976	1.3822	0.3301	0.1735
1981	1.4322	0.3655	0.1932
1986	1.5183	0.4212	0.2253

Table B.8 : Measures of Collaboration for Computer Science

Year	CI	DC	CC
1961	1.3369	0.2521	0.1384
1966	1.4660	0.3395	0.1874
1971	1.5096	0.3515	0.1968
1976	1.6267	0.4412	0.2458
1981	1.7484	0.4936	0.2801
1986	1.8226	0.5675	0.3191

Table B.9 : Measures of Collaboration for Sociology

Year	CI	DC	CC
1961	1.2899	0.2210	0.1201
1966	1.2276	0.1776	0.0959
1971	1.2591	0.2019	0.1092
1976	1.3503	0.2670	0.1456
1981	1.3340	0.2465	0.1356
1986	1.4143	0.2877	0.1607

Table B.10 : Measures of Collaboration for Political Science

Year	CI	DC	CC
1961	1.0440	0.0379	0.0198
1966	1.0780	0.0672	0.0353
1971	1.1023	0.0875	0.0460
1976	1.1191	0.0980	0.0521
1981	1.1398	0.1164	0.0616
1986	1.1618	0.1312	0.0699

Table B.11 : Measures of Collaboration for Economic Science and Managerial Sciences

Year	CI	DC	CC
1961	1.0990	0.0890	0.0462
1966	1.1187	0.1088	0.0560
1971	1.1771	0.1562	0.0816
1976	1.3204	0.2680	0.1422
1981	1.4017	0.3249	0.1744
1986	1.4469	0.3597	0.1917

Table B.12 : Measures of Collaboration for Psychology

Year	CI	DC	CC
1961	1.5354	0.3755	0.2103
1966	1.6441	0.4406	0.2486
1971	1.7106	0.4600	0.2643
1976	1.8294	0.5315	0.3060
1981	1.9666	0.5652	0.3331
1986	2.0663	0.5717	0.3469

Table B.13 : Measures of Collaboration for Library and Information Science

Year	CI	DC	CC
1961	1.0660	0.0590	0.0306
1966	1.0748	0.0570	0.0311
1971	1.1880	0.1365	0.0752
1976	1.1778	0.1294	0.0711
1981	1.2224	0.1653	0.0904
1986	1.2356	0.1712	0.0938

Table B.14 : Measures of Collaboration for History

Year	CI	DC	CC
1961	1.0294	0.0242	0.0128
1966	1.0551	0.0454	0.0240
1971	1.0365	0.0324	0.0167
1976	1.0553	0.0480	0.0250
1981	1.0651	0.0585	0.0303
1986	1.0755	0.0642	0.0337

Table B.15 : Measures of Collaboration for English Studies

Year	CI	DC	CC
1961	1.0165	0.0150	0.0077
1966	1.0154	0.0146	0.0075
1971	1.0183	0.0174	0.0088
1976	1.0233	0.0221	0.0112
1981	1.0237	0.0216	0.0111
1986	1.0241	0.0226	0.0115

Table B.16 : Measures of Collaboration for Philosophy

Year	CI	DC	CC
1967/68	1.0344	0.0313	0.0162
1969	1.0294	0.0289	0.0145
1971	1.0346	0.0313	0.0162
1976	1.0277	0.0261	0.0133
1981	1.0457	0.0415	0.0214
1986	1.0600	0.0541	0.0280

APPENDIX C : Forms of the Discrete Probability Models

In the following, we write

$f(x)$ = frequency of occurrence of x

n = sample size

MLE = maximum likelihood estimate

x = sample mean

s^2 = sample variance

1. Zipf

$$p(x) = k\, x^{-b}$$

where $0 < k < 1$, $b > 0$ for $x = 1,2,....$

The maximum likelihood estimator of b satisfies the equation

$$f(x) \ln x \,/\, \Sigma f(x) = -\zeta(b) \,/\, \zeta(b) ,$$

where $\zeta(.)$ denotes the Riemann zeta function ;

and

$$k = [\Sigma\, x^{-b}]^{-1}$$

2. Mandelbrot

$$p(x) = \frac{k}{n} (x+c)^{-b} ;$$

where $k > 0$, $b > 0$, $-1 < c < \infty$ for $x = 1,2,...$

The parameters are estimated by non-linear least-squares method (Schreiner, Kramer, Krischer & Langsam, 1985).

3. Geometric

$$p(x) = p\,(1-p)^{x-1} ;$$

where $0 < p < 1$ for $x = 1,2,...$

The MLE of p is $\hat{p} = 1/\bar{x}$

4. Shifted Poisson

$$p(x) = \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \quad ;$$

where $\lambda > 0$ for $x = 1,2,\ldots$

MLE of λ is $\hat{\lambda} = \bar{x} - 1$

5. Shifted Generalized Poisson

$$p(x) = \frac{\lambda_1(\lambda_1 + (x-1)\lambda_2)^{x-2}\, e^{-(\lambda_1 + (x-1)\lambda_2)}}{(x-1)!} \quad ;$$

where $\lambda_1 > 0$, $|\lambda_2| < 1$ for $x = 1,2,\ldots$

The moment estimates of λ_1 and λ_2 are

$$\hat{\lambda}_2 = 1 - ((\bar{x}-1)/s^2)^{\frac{1}{2}} \quad \text{and}$$

$$\hat{\lambda}_1 = (\bar{x}-1)(1 - \hat{\lambda}_2)$$

The maximum likelihood estimator of λ_2 is given by the root of the equation

$$\sum_{x=1}^{x_{max}} \frac{f(x)(x-1)(x-2)}{(\bar{x}-1 + (x-\bar{x})\hat{\lambda}_2)} - n(\bar{x}-1) = 0$$

and

$$\hat{\lambda}_1 = (\bar{x}-1)(1 - \hat{\lambda}_2)$$

6. Logarithmic Series

$$p(x) = \frac{\alpha\theta^x}{x} \quad ;$$

where $0 < \theta < 1$, $\alpha = -[\ln(1-\theta)]^{-1}$ for $x = 1,2,\ldots$

The moment estimate of θ is

$$\hat{\theta} = 1 - \frac{\bar{x}}{(s^2 + \bar{x}^2)}$$

The maximum likelihood estimator of θ is given by the root of the equation

$$\bar{x} + \frac{\hat{\theta}}{(1 - \hat{\theta}) \ln (1 - \hat{\theta})} = 0 ;$$

7. Borel - Tanner

$$p(x) = \frac{x^{x-2} \alpha^{x-1} e^{-\alpha x}}{(x-1)!} ;$$

where $\alpha > 0$ for $x = 1,2,...$

The moment estimate of α is given by

$$\hat{\alpha} = 1 - 1/\bar{x}$$

8. Shifted Yule

$$p(x) = \frac{\alpha \Gamma(x) \Gamma(\alpha+1)}{\Gamma(\alpha+x+1)} ;$$

where $\alpha > 0$ for $x = 1,2,...$

The moment estimate of α is given by

$$\hat{\alpha} = 1 + 1/(\bar{x}-1) = \bar{x}/(\bar{x}-1)$$

9. Shifted Waring

$$p(x) = \frac{\alpha}{\alpha + \beta} \frac{\Gamma(\alpha+\beta+1) \Gamma(x+\beta-1)}{\Gamma(\alpha+\beta+x)\Gamma(\beta)} ;$$

where $\alpha > 0$, $\beta > 0$ for $x = 1,2,...$

The moment estimates of α and β are given by

$$\hat{\alpha} = \frac{2s^2}{(s^2 - \bar{x}(\bar{x}-1))} ; \text{ and}$$

$$\hat{\beta} = (\hat{\alpha}-1)(\bar{x}-1)$$

Note that the mean exists only if $\alpha > 1$ while the variance exists only if $\alpha > 2$.

The maximum likelihood estimator of α is given by the root of the equation

$$\frac{n(\bar{x}-1)(\hat{\alpha}-1)}{\hat{\alpha}[\bar{x}(\hat{\alpha}-1)+1]} - \sum_{x=2}^{x_{max}} f(k) \sum_{x=2}^{k} \frac{1}{[\bar{x}(\hat{\alpha}-1)+x]} = 0 \; ; \quad \text{and}$$

$$\hat{\beta} = (\hat{\alpha}-1)(\bar{x}-1)$$

10. Shifted Generalized Waring

$$p(x) = \frac{\Gamma(v+\alpha)\ \Gamma(x+v-1)\ \Gamma(x+\beta-1)}{B(\alpha,\beta)\ \Gamma(v)\ \Gamma(x+v+\alpha+\beta-1)\ (x-1)!} \; ;$$

where $\alpha > 0$, $\beta > 0$, $v > 0$ for $x = 1,2,...$

The three parameters can be estmated from the first three ascending factorial moments given by

$$\mu_{[1]} = \frac{\beta v}{\alpha-1} \; ; \quad \alpha > 1$$

$$\mu_{[2]} = \frac{\beta v\ (v+1)(\beta+1)}{(\alpha-1)(\alpha-2)} \; ; \quad \alpha > 2 \quad \text{and}$$

$$\mu_{[3]} = \frac{\beta v\ (v+1)(v+2)(\beta+1)(\beta+2)}{(\alpha-1)(\alpha-2)(\alpha-3)} \; ; \quad \alpha > 3$$

11. Shifted Inverse Gaussian-Poisson

$$p(x) = (2\alpha/\pi)^{\frac{1}{2}}\ e^{\alpha(1-\theta)^{1/2}} \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} K_{x-3/2}(\alpha) \; ;$$

where $0 \le \theta \le 1$, $\alpha \ge 0$, $K_v(z)$ is the modified Bessel function of the second kind of order v with argument z for $x = 1,2,...$

Estimation of parameters :

(a) If the observed frequency distribution is reversed J-shape, with a large proportionate frequency in the unit cell.

$$\hat{\theta} = 1 - \left[\frac{-\ln p(x=1)}{2(x-1) + \ln p(x=1)} \right]^2 \; ; \text{ and}$$

$$\hat{\alpha} = \frac{2(\bar{x}-1)}{\hat{\theta}} (1 - \hat{\theta})^{\frac{1}{2}}$$

(b) If observed frequency distribution is unimodal.

The moment estimates of θ and α are given by

$$\hat{\theta} = 1 - \left[\frac{2s^2}{(\bar{x}-1)} - 1 \right]^{-1} \; ; \text{ and}$$

$$\hat{\alpha} = \frac{2(\bar{x}-1)}{\hat{\theta}} (1 - \hat{\theta})^{\frac{1}{2}}$$

The maximum likelihood estimator of α is given by the root of the equation

$$[\hat{\alpha}(\epsilon^2+\hat{\alpha}^2)^{-\frac{1}{2}}][1 + (\bar{x}-1)/w] - (1/n) \sum_{x=1}^{max} f(x) \, R_{x-3/2}(\hat{\alpha}) = 0$$

where $\epsilon = \bar{x} - 1$, $w = (\epsilon^2+\hat{\alpha}^2)^{\frac{1}{2}} - \epsilon$, $R_v(z) = \dfrac{K_{v+1}(z)}{K_v(z)}$

and

$$\theta = \frac{-2\epsilon^2 \pm 2\epsilon(\epsilon^2+\hat{\alpha}^2)^{\frac{1}{2}}}{\hat{\alpha}^2}$$

12. Shifted Generalized Inverse Gaussian-Poisson

$$p(x) = \frac{(1-\theta)^{\tau/2}}{K_\tau[\alpha(1-\theta)^{\frac{1}{2}}]} \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} K_{x-1+\tau}(\alpha) \quad ;$$

where $-\infty < \tau < \infty$, $0 \le \theta < 1$, $\alpha > 0$, $K_v(z)$ is the modified Bessel function of the second kind of order v with argument z for $x = 1,2,...$

If we substitute $\epsilon = \alpha\theta / [2(1-\theta)^{\frac{1}{2}}]$, we obtain

$$p(x) = \frac{(w/\alpha)^{\tau} \ (\epsilon w/\alpha)^{x-1}}{K_{\tau}(w) \ (x-1)!} \quad K_{x-1+\tau}(\alpha) \quad ; \quad x = 1,2,...$$

where $w = (\epsilon^2+\alpha^2)^{\frac{1}{2}} - \epsilon$

For fixed τ , the likelihood equations are

$$\hat{\alpha}(\hat{\epsilon}^2+\hat{\alpha}^2)^{\frac{1}{2}} \ [R_{\tau}(w) + (\bar{x}-1)/w] - (1/n) \sum_{x=1}^{x_{max}} R_{x-1+\tau}(\hat{\alpha}) \ f(x) = 0$$

and

$$\hat{\epsilon}R_{\tau}(w) = \bar{x} - 1$$

These equations can be solved using Newton-Ralphson iteration method while τ can be estimated using Stein, Zucchini and Juritz's algorithm (1987).

13. Shifted Binomial

$$p(x) = \binom{v}{x-1} p^{x-1} \quad (1-p)^{v-(x-1)} \quad ;$$

where $0 < p < 1$, $v > 0$ for $x = 1,2, ..,v,v+1$

The moment estimates of p and v are given by

$$\hat{p} = 1 - \frac{s^2}{\bar{x}-1} \quad ; \text{ and}$$

$$\hat{v} = \frac{\bar{x}-1}{\hat{p}}$$

The maximum likelihood estimator of v is given by the root of the equation

$$\sum_{j=1}^{x_{max}-1} (\hat{v}-j+1)^{-1} f_j + n \ln [1 - (\bar{x}-1)/\hat{v}] = 0 \ ;$$

where f_j = number of x's which exceed j ; and

$$\hat{p} = \frac{\bar{x}-1}{\hat{v}}$$

14. Shifted Beta-Binomial

$$p(x) = \binom{v}{x-1} \frac{\Gamma(\alpha+x-1)\ \Gamma(v+\beta-x+1)\ \Gamma(\alpha+\beta)}{\Gamma(v+\alpha+\beta)\ \Gamma(\alpha)\ \Gamma(\beta)} \ ; \ x = 1,2,..,v+1$$

where $\alpha,\ \beta > 0$

Estimation of Parameters :

1) If there is a large proportion of unit values, then β is given by the root of the equation

$$p(1) - \frac{\Gamma(v+\hat{\beta})\ \Gamma[\hat{\beta}v/(v-\bar{x}+1)]}{\Gamma[v+(\hat{\beta}v/(v-\bar{x}+1))]\ \Gamma(\hat{\beta})} = 0 \quad ; \quad \text{and}$$

$$\hat{\alpha} = (\bar{x}-1)\hat{\beta}/(v-\bar{x}+1)$$

2) Otherwise, the moment estimates of α and β are given by

$$\hat{\alpha} = \frac{(\bar{x}-1)\ [(\bar{x}-1)v - (\bar{x}-1)^2 - s^2]}{[vs^2 - (\bar{x}-1)v + (\bar{x}-1)^2]} \quad ; \quad \text{and}$$

$$\hat{\beta} = \hat{\alpha}(v-\bar{x}+1)/(\bar{x}-1)$$

15. Shifted Negative Binomial

$$p(x) = \binom{v+x-2}{x-1} p^v\ (1-p)^{x-1} \quad ;$$

where $0 < p < 1,\ v > 0$ for $x = 1,2,...$

The momemt estimates of p and v are given by

$$\hat{p} = \frac{\bar{x}-1}{s^2} \quad ; \text{ and}$$

$$\hat{v} = \frac{\hat{p}(\bar{x}-1)}{1-\hat{p}}$$

The maximum likelihood estimator of v is given by the root of the equation

$$\ln\ [1 + (\bar{x}-1)/v] - \sum_{j=2}^{x_{max}} \frac{F_j}{(v+j-2)} = 0 \quad ,$$

where Fj = proportion of x's which are greater than or equal to j; and

$$\hat{p} = \frac{\hat{v}}{(\bar{x}+\hat{v}-1)}$$

16. Shifted Generalized Negative Binomial

$$p(x) = \frac{v\,\Gamma[v+\beta(x-1)]}{(x-1)!\;\Gamma[v+\beta(x-1)-(x-1)+1]}\,\alpha^{x-1}\,(1-\alpha)^{v+\beta(x-1)-(x-1)} \;;$$

where $0 < \alpha < 1$, $|\alpha\beta| < 1$, $v > 0$ for $x = 1,2,\ldots$

The moment esimates of α, β, and v are given by

$$\hat{\alpha} = 1 - \tfrac{1}{2}A + (\tfrac{1}{4}A^2 - 1)^{\frac{1}{2}} \;;$$

$$\hat{\beta} = 1/\hat{\alpha}\,[1 - ((\bar{x}-1)(1-\hat{\alpha})/\mu_2)^{\frac{1}{2}}] \;; \text{ and}$$

$$\hat{v} = (\bar{x}-1)(1-\hat{\alpha}\hat{\beta})/\hat{\alpha}$$

where $A = -2 + \frac{[(\bar{x}-1)\mu_3 - 3\mu_2^2]^2}{(\bar{x}-1)\mu_2^3}$

and μ_2 and μ_3 are the second and third central moments respectively.

APPENDIX D : RESULTS OF THE CHI SQUARE TESTS

Table D.1 : Results of the Chi Square Goodness-of-fit Test for Physics

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	0	0	1
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	2	0	0	0
Sh. Gen. Waring	6	0	0	0
Shifted IGP	6	1	1	5
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	2	0	0	0
Sh. Neg. Binomial	6	0	0	0
Sh. Gen. Neg. Bin.	1	0	0	0

Table D.2 : Results of the Chi Square Goodness-of-fit Test for Chemical Sciences

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	0	0	1
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	6	0	0	0
Shifted IGP	6	0	1	5
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	2	0	0	0
Sh. Neg. Binomial	6	0	0	0
Sh. Gen. Neg. Bin.	0	0	0	0

Table D.3 : Results of the Chi Square Goodness-of-fit Test for Biological Sciences

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	1	2	0
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	6	0	0	0
Shifted IGP	6	2	3	5
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	3	0	0	0
Sh. Neg. Binomial	6	1	1	1
Sh. Gen. Neg. Bin.	0	0	0	0

Table D.4 : Results of the Chi Square Goodness-of-fit Test for Mathematics

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	1	1	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	1	1	0
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	0	0	0	0
Shifted IGP	6	1	1	6
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	0	0	0	0
Sh. Neg. Binomial	6	0	1	0
Sh. Gen. Neg. Bin.	0	0	0	0

Table D.5 : Results of the Chi Square Goodness-of-fit Test for Statistics

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	2	2	0
Shifted Poisson	6	2	4	0
Sh. Gen. Poisson	6	5	5	2
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	0	0	0	0
Shifted IGP	2	2	2	2
Shifted GIGP	0	0	0	0
Shifted Binomial	4	3	3	0
Sh. Beta-Binomial	2	2	2	0
Sh. Neg. Binomial	2	2	2	0
Sh. Gen. Neg. Bin.	3	3	3	2

Table D.6 : Results of the Chi Square Goodness-of-fit Test for Computer Science

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	1	2	1
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	1	3	0
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	4	0	0	0
Shifted IGP	6	2	3	5
Shifted GIGP	1	0	1	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	4	1	1	0
Sh. Neg. Binomial	6	1	2	0
Sh. Gen. Neg. Bin.	2	1	2	0

Table D.7 : Results of the Chi Square Goodness-of-fit Test for Sociology

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	2	3	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	5	6	1
Logarithmic	6	1	1	0
Borel-Tanner	6	0	1	0
Shifted Yule	6	0	0	0
Shifted Waring	5	4	5	2
Sh. Gen. Waring	6	0	0	0
Shifted IGP	6	5	6	3
Shifted GIGP	2	1	2	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	0	1	0
Sh. Neg. Binomial	6	4	5	0
Sh. Gen. Neg. Bin.	2	1	2	0

Table D.8 : Results of the Chi Square Goodness-of-fit Test for Political Science

Model	No. of Valid Results	No. of Passes 5%	1%	No. of Best Fits Provided
Zipf	6	2	2	0
Mandelbrot	6	4	4	0
Geometric	6	0	1	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	5	5	0
Logarithmic	6	4	4	0
Borel-Tanner	6	5	5	0
Shifted Yule	6	4	4	1
Shifted Waring	6	5	6	0
Sh. Gen. Waring	0	0	0	0
Shifted IGP	6	5	6	4
Shifted GIGP	3	2	2	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	0	0	0
Sh. Neg. Binomial	6	5	5	1
Sh. Gen. Neg. Bin.	5	3	4	0

Table D.9 : Results of the Chi Square Goodness-of-fit Test for Economic Science and Managerial Sciences

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	1	2	0
Geometric	6	3	3	0
Shifted Poisson	6	3	4	1
Sh. Gen. Poisson	6	4	6	0
Logarithmic	6	1	3	0
Borel-Tanner	6	1	1	0
Shifted Yule	6	1	1	0
Shifted Waring	1	1	1	0
Sh. Gen. Waring	1	0	0	0
Shifted IGP	6	5	6	3
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	3	4	0
Sh. Neg. Binomial	6	4	6	0
Sh. Gen. Neg. Bin.	5	4	5	2

Table D.10 : Results of the Chi Square Goodness-of-fit Test for Psychology

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	0	0	0
Mandelbrot	6	0	0	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	4	5	1
Logarithmic	6	0	0	0
Borel-Tanner	6	0	0	0
Shifted Yule	6	0	0	0
Shifted Waring	0	0	0	0
Sh. Gen. Waring	6	1	2	0
Shifted IGP	6	4	6	3
Shifted GIGP	0	0	0	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	1	1	0
Sh. Neg. Binomial	6	4	4	2
Sh. Gen. Neg. Bin.	2	2	2	0

Table D.11 : Results of the Chi Square Goodness-of-fit Test for Library and Information Science

Model	No. of Valid Results	No. of Passes 5%	1%	No. of Best Fits Provided
Zipf	6	2	2	0
Mandelbrot	6	4	5	1
Geometric	6	1	1	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	5	5	0
Logarithmic	6	1	1	0
Borel-Tanner	6	2	3	0
Shifted Yule	6	3	5	1
Shifted Waring	6	6	6	0
Sh. Gen. Waring	4	0	0	0
Shifted IGP	6	6	6	3
Shifted GIGP	3	3	3	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	0	0	0
Sh. Neg. Binomial	6	4	5	1
Sh. Gen. Neg. Bin.	5	4	4	0

Table D.12 : Results of the Chi Square Goodness-of-fit Test for History

Model	No. of Valid Results	No. of Passes 5%	1%	No. of Best Fits Provided
Zipf	6	1	2	1
Mandelbrot	6	3	4	0
Geometric	6	0	0	0
Shifted Poisson	6	0	0	0
Sh. Gen. Poisson	6	5	5	0
Logarithmic	6	0	0	0
Borel-Tanner	6	1	2	0
Shifted Yule	6	1	2	0
Shifted Waring	6	5	6	0
Sh. Gen. Waring	3	0	0	0
Shifted IGP	6	5	6	5
Shifted GIGP	1	1	1	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	0	0	0
Sh. Neg. Binomial	6	3	5	0
Sh. Gen. Neg. Bin.	3	2	3	0

Table D.13 : Results of the Chi Square Goodness-of-fit Test for English Studies

Model	No. of Valid Results	No. of Passes 5%	1%	No. of Best Fits Provided
Zipf	6	5	6	0
Mandelbrot	6	6	6	0
Geometric	6	5	6	0
Shifted Poisson	6	6	6	0
Sh. Gen. Poisson	6	6	6	1
Logarithmic	6	5	5	0
Borel-Tanner	6	4	5	0
Shifted Yule	6	5	5	3
Shifted Waring	6	6	6	0
Sh. Gen. Waring	0	0	0	0
Shifted IGP	6	6	6	1
Shifted GIGP	1	1	1	0
Shifted Binomial	0	0	0	0
Sh. Beta-Binomial	6	0	0	0
Sh. Neg. Binomial	6	6	6	0
Sh. Gen. Neg. Bin.	5	5	5	1

Table D.14 : Results of the Chi Square Goodness-of-fit Test for Philosophy

Model	No. of Valid Results	No. of Passes 5%	No. of Passes 1%	No. of Best Fits Provided
Zipf	6	3	4	0
Mandelbrot	6	5	6	4
Geometric	6	1	4	0
Shifted Poisson	6	1	1	1
Sh. Gen. Poisson	6	6	6	0
Logarithmic	6	2	2	0
Borel-Tanner	6	2	2	0
Shifted Yule	6	4	6	0
Shifted Waring	5	3	5	0
Sh. Gen. Waring	0	0	0	0
Shifted IGP	5	5	5	0
Shifted GIGP	0	0	0	0
Shifted Binomial	1	1	1	0
Sh. Beta-Binomial	5	0	0	0
Sh. Neg. Binomial	5	5	5	0
Sh. Gen. Neg. Bin.	6	6	6	1

APPENDIX E : FITS OF THE SHIFTED INVERSE GAUSSIAN POISSON DISTRIBUTION TO 1986 DATA SETS

Table E.1 : Fit of IGP to 1986 Engineering Sciences Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	3050	3022.9
2	2500	2654.8
3	1682	1512.6
4	792	737.3
5	274	338.6
6	130	152.7
7	58	68.7
8	21	31.1
9	10	14.2
10	7	6.5
11	6	3.0
12-16	14	2.6
Total	8544	
Mean	2.27	
Variance	2.07	
α		3.35
θ		0.52
X^2		106.33
d.f.		9
$p(X^2)$		0.82E-18

Table E.2 : Fit of IGP to 1986 Medical Sciences Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	1204	1049.3
2	1252	1450.8
3	1123	1227.4
4	877	842.1
5	652	521.2
6	330	306.3
7	166	175.4
8	86	99.2
9	54	55.9
10	24	31.5
11	12	17.7
12	10	10.0
13	6	5.7
14-16	4	6.1
Total	5800	
Mean	3.24	
Variance	3.94	
α		4.47
θ		0.62
X^2		101.73
d.f.		11
$p(X^2)$		0.81E-16

Table E.3 : Fit of IGP to 1986 Physics Data Set

Number of Authors	Number of Papers Observed	Expected
1	1066	1182.1
2	1328	1224.7
3	944	840.7
4	535	502.3
5	249	286.8
6	128	162.2
7	62	92.2
8	32	53.0
9	22	30.8
10	11	18.1
11	8	10.7
12	10	6.4
13	6	3.9
14	4	2.4
15-37	15	3.8
Total	4420	
Mean	2.81	
Variance	5.17	
α		3.08
θ		0.67
X^2		109.03
d.f.		12
$p(X^2)$		0.93E-17

Table E.4 : Fit of IGP to 1986 Chemical Sciences Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	1818	2131.1
2	2312	2252.6
3	1932	1559.4
4	1061	930.3
5	475	526.1
6	223	293.3
7	93	163.7
8	46	92.1
9	25	52.3
10	17	30.0
11	14	17.3
12	15	10.1
13	10	6.0
14-30	31	8.8
Total	8072	
Mean	2.80	
Variance	3.51	
α		3.23
θ		0.65
X^2		312.09
d.f.		11
$p(X^2)$		< E-20

Table E.5 : Fit of IGP to 1986 Biological Sciences Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	1367	1565.9
2	2643	2429.5
3	2303	2200.5
4	1559	1546.8
5	900	944.1
6	483	529.8
7	247	282.4
8	119	145.7
9	60	73.7
10	35	36.8
11	25	18.2
12	25	9.0
13	8	4.4
14-18	17	4.2
Total	9791	
Mean	3.24	
Variance	3.59	
α		5.97
θ		0.52
X^2		140.68
d.f.		11
$p(X^2)$		< E-20

Table E.6 : Fit of IGP to 1986 Mathematics Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	25535	25695.5
2	10399	10074.3
3	2313	2405.4
4	385	463.7
5	70	80.4
6	16	13.2
7-16	17	2.5
Total	38735	
Mean	1.43	
Variance	0.48	
α		4.59
θ		0.17
X^2		114.72
d.f.		4
$p(X^2)$		< E-20

Table E.7 : Fit of IGP to 1986 Computer Science Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	1263	1322.2
2	1123	1010.4
3	403	420.7
4	85	127.2
5	25	31.4
6	14	6.7
7-9	7	1.6
Total	2920	
Mean	1.82	
Variance	0.89	
α		11.16
θ		0.14
X^2		57.80
d.f.		4
$p(X^2)$		0.84E-11

Table E.8 : Fit of IGP to 1986 Sociology Data Set

Number of Authors	Number of Papers Observed	Number of Papers Expected
1	6856	6854.8
2	1962	1970.3
3	555	538.7
4	161	166.9
5	57	57.8
6	17	21.7
7	9	8.6
8-12	8	6.1
Total	9625	
Mean	1.41	
Variance	0.64	
α		1.11
θ		0.52
X^2		2.35
d.f.		5
$p(X^2)$		0.80

Table E.9 : Fit of IGP to 1986 Political Science Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	5191	5192.4
2	648	643.6
3	106	106.4
4	22	23.6
5	4	6.2
6-11	4	2.7
Total	5975	
Mean	1.16	
Variance	0.23	
α		0.60
θ		0.41
X^2		1.61
d.f.		3
$p(X^2)$		0.66

Table E.10 : Fit of IGP to 1986 Economic Science and Managerial Sciences Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	235	240.1
2	110	96.9
3	16	24.0
4-6	6	5.9
Total	367	
Mean	1.45	
Variance	0.50	
α		4.39
θ		0.18
X^2		4.55
d.f.		1
$p(X^2)$		0.03

Table E.11 : Fit of IGP to 1986 Psychology Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	2185	2142.1
2	1474	1566.8
3	806	780.4
4	374	346.3
5	156	149.4
6	66	64.6
7	20	28.3
8	10	12.6
9	4	5.7
10-11	6	3.8
Total	5101	
Mean	2.07	
Variance	1.64	
α		2.76
θ		0.53
X^2		14.50
d.f.		7
$p(X^2)$		0.04

Table E.12 : Fit of IGP to 1986 Library and Information Science Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	4971	4971.1
2	786	776.2
3	170	169.4
4	36	50.6
5	17	18.1
6	10	7.2
7	3	3.0
8-10	5	2.2
Total	5998	
Mean	1.24	
Variance	0.41	
α		0.56
θ		0.56
X^2		8.98
d.f.		5
$p(X^2)$		0.11

Table E.13 : Fit of IGP to 1986 History Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	17019	17022.5
2	1021	1001.1
3	104	127.9
4	34	25.7
5	4	6.4
6-10	4	2.5
Total	2516	
Mean	1.03	
Variance	0.04	
α		0.30
θ		0.39
X^2		9.29
d.f.		3
$p(X^2)$		0.03

Table E.14 : Fit of IGP to 1986 English Studies Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	2253	2253.0
2	55	55.4
3	4	3.2
Total	2612	
Mean	1.02	
Variance	0.03	
α		0.23
θ		0.19
X^2		0.18
d.f.		0*
$p(X^2)$		

* This is a case of overfitting

Table E.15 : Fit of IGP to 1986 Philosophy Data Set

	Number of Papers	
Number of Authors	Observed	Expected
1	7079	7079.0
2	361	367.2
3	44	32.7
Total	7484	
Mean	1.06	
Variance	0.07	
α		0.41
θ		0.25
X^2		4.00
d.f.		0*
$p(X^2)$		

* This is a case of overfitting

APPENDIX F : DERIVATION OF THE SHIFTED INVERSE GAUSSIAN-POISSON AS A MODEL FOR THE DISTRIBUTION OF AUTHORSHIPS

Let the average number of co-authors working with a principal researcher on a paper be λ_i. Further, make the reasonable assumption that the statistical variability around this average λ_i follows a Poisson distribution. In a group of authors, the λ_i's themselves would vary considerably as some authors like to collaborate a lot while others prefer to work alone but only collaborate if deemed extremely necessary. If we assume now that the λ_i's themselves are distributed according to the inverse Gaussian distribution law, we can then obtain the corresponding compound distribution, i.e., the inverse Gaussian-Poisson distribution, as shown by Sichel (1971).

Hence, the distribution of the number of co-authors, r, working with a principal researcher on a paper follows an inverse Gaussian-Poisson distribution.

$$\therefore \quad p(r) = (2\alpha/\pi)^{\frac{1}{2}} e^{\alpha(1-\theta)^{\frac{1}{2}}} \frac{(\frac{1}{2}\alpha\theta)^r}{r!} K_{r-\frac{1}{2}}(\alpha) \; ; \; r = 0,1,\ldots$$

Now, let x = number of authors

= number of co-authors + principal researcher

= r + 1

$\Longrightarrow$ r = x - 1

If we substitute r = x - 1 in the above equation, we obtain

$$p(r=x-1) = (2\alpha/\pi)^{\frac{1}{2}} e^{\alpha(1-\theta)^{\frac{1}{2}}} \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} K_{x-3/2}(\alpha) \; ; \; x = 1,2,...$$

Now, the probability of a paper without an author is zero, i.e., $p(x=0) = 0$. Also, the probability of a paper having x authors is the same as the probability of a principal researcher working with (x-1) co-authors, i.e., $p(x) = p(r=x-1)$.

$$\therefore \; p(x) = p(r=x-1)$$

$$= (2\alpha/\pi)^{\frac{1}{2}} e^{\alpha(1-\theta)^{\frac{1}{2}}} \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} K_{x-3/2}(\alpha) \; ; \; x = 1,2,..$$

Hence, the number of authors of a paper follows a shifted inverse Gaussian-Poisson distribution.

APPENDIX F : DERIVATION OF E(1/X)

(i) Shifted Generalized Poisson

$$p(x) = \frac{\lambda_1 [\lambda_1 + (x-1)\lambda_2]^{x-2} \; e^{-[\lambda_1 + (x-1)\lambda_2]}}{(x-1)!} \qquad ; \; x = 1,2,\ldots$$

$$= \frac{\lambda_1 [1 + (x-1)\frac{\lambda_2}{\lambda_1}]^{x-2} \; \lambda_1^{(x-2)} \; e^{-[1 + (x-1)\frac{\lambda_2}{\lambda_1}]\lambda_1}}{(x-1)!}$$

$$= \frac{\lambda_1^{x-1} [1 + (x-1)\frac{\lambda_2}{\lambda_1}]^{x-2} \; e^{-\lambda_1} \; e^{-[(x-1)\frac{\lambda_2}{\lambda_1}]\lambda_1}}{(x-1)!} \quad ;$$

If we set $\frac{\lambda_2}{\lambda_1} = \alpha$, we obtain

$$p(x) = \frac{\lambda_1^{x-1} [1 + (x-1)\alpha]^{x-2} \; e^{-[(x-1)\alpha]\lambda_1}}{(x-1)! \; e^{\lambda_1}}$$

$$= \frac{\lambda_1^{x-1} [1 + (x-1)\alpha]^{x-2} \; [e^{-\alpha\lambda_1}]^{x-1}}{(x-1)! \; e^{\lambda_1}}$$

$$= \frac{[1 + (x-1)\alpha]^{x-2} \; [\lambda_1 e^{-\alpha\lambda_1}]^{x-1}}{(x-1)! \; e^{\lambda_1}}$$

$$= \frac{[1 + (x-1)\alpha]^{x-2}}{(x-1)!} \quad \frac{(\lambda_1 e^{-\alpha\lambda_1})^{x}}{\lambda_1 e^{\lambda_1} \; e^{-\alpha\lambda_1}}$$

$$= \frac{[1 + (x-1)\alpha]^{x-2}}{(x-1)!} \quad \frac{(\lambda_1 e^{-\alpha\lambda_1})^{x}}{\lambda_1 e^{\lambda_1(1-\alpha)}}$$

$$= a(x) \quad \frac{[g(\lambda_1)]^{x}}{f(\lambda_1)}$$

where, $a(x) = \frac{[1 + (x-1)\alpha]^{x-2}}{(x-1)!}$;

$f(\lambda_1) = \lambda_1 e^{\lambda_1(1-\alpha)}$;

$g(\lambda_1) = \lambda_1 e^{-\alpha\lambda_1}$

According to Kumar and Consul (1979), the recursive relation between the negative moments about the point -k for modified power series distribution is

$$M(r,k) = [f(\lambda_1)\ (g(\lambda_1))^k]^{-1} \int_0^{\lambda_1} M(r-1,k)\ I(k)\ d\lambda_1;\ r = 1,2,\ldots$$

where, $I(k) = f(\lambda_1)\ g'(\lambda_1)\ [g(\lambda_1)]^{k-1}$; $M(0,k) = 1$

Now,

$$E[1/X] = M(1,0) = [f(\lambda_1)]^{-1} \int_0^{\lambda_1} f(\lambda_1)\ g'(\lambda_1)\ [g(\lambda_1)]^{-1}\ d\lambda_1$$

$$= \frac{1}{\lambda_1 e^{\lambda_1(1-\alpha)}} \int_0^{\lambda_1} \frac{\lambda_1 e^{\lambda_1(1-\alpha)}(1-\alpha\lambda_1)\ e^{-\alpha\lambda_1}}{\lambda_1 e^{-\alpha\lambda_1}}\ d\lambda_1$$

$$= \frac{1}{\lambda_1 e^{\lambda_1(1-\alpha)}} \int_0^{\lambda_1} e^{\lambda_1(1-\alpha)}\ (1-\alpha\lambda_1)\ d\lambda_1$$

$$= \frac{1}{\lambda_1 e^{\lambda_1(1-\alpha)}} \int_0^{\lambda_1} [e^{\lambda_1(1-\alpha)} - \alpha\lambda_1 e^{\lambda_1(1-\alpha)}]\ d\lambda_1$$

$$= \frac{e^{\lambda_1(1-\alpha)} - 1 - \alpha\lambda_1 e^{\lambda_1(1-\alpha)} + \alpha^2\lambda_1 e^{\lambda_1(1-\alpha)}}{\lambda_1(1-\alpha)^2\ e^{\lambda_1(1-\alpha)}}$$

Now, substituting back $\alpha = \frac{\lambda_2}{\lambda_1}$, we have

$$E[1/X] = \frac{\lambda_1 e^{\lambda_1-\lambda_2} - \lambda_1 - \lambda_1\lambda_2 e^{\lambda_1-\lambda_2} + \lambda_2^2\ e^{\lambda_1-\lambda_2}}{(\lambda_1-\lambda_2)^2\ \ e^{\lambda_1-\lambda_2}}$$

(ii) Shifted Inverse Gaussian-Poisson

For the shifted generalized inverse Gaussian-Poisson,

$$p(x) = \frac{(1-\theta)^{\tau/2}}{K_\tau[\alpha(1-\theta)^{\frac{1}{2}}]} \quad \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} \quad K_{x-1+\tau}(\alpha)$$

$$= C \quad \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} \quad K_{(x-1)+\tau}(\alpha)$$

$$E[1/X] = C \sum_{x=1} (1/x) \quad \frac{(\frac{1}{2}\alpha\theta)^{x-1}}{(x-1)!} \quad K_{[(x-1)+\tau]}(\alpha)$$

$$= \frac{2C}{\alpha\theta} \sum_{x=1} \frac{(\frac{1}{2}\alpha\theta)^{x}}{x!} \quad K_{[(x-1)+\tau]}(\alpha)$$

$$= \frac{2C}{\alpha\theta} \sum_{x=1} \frac{(\frac{1}{2}\alpha\theta)^{x}}{x!} \quad K_{[x+(\tau-1)]}(\alpha)$$

but $$\frac{(1-\theta)^{(\tau-1)/2}}{K_{\tau-1}[\alpha(1-\theta)^{\frac{1}{2}}]} \sum_{x=0} \frac{(\frac{1}{2}\alpha\theta)^{x}}{x!} \quad K_{x+\tau-1}(\alpha) = 1$$

$$E[1/X] = \frac{2C}{\alpha\theta} [1 - K_{\tau-1}(\alpha)] \quad \frac{K_{\tau-1}[\alpha(1-\theta)^{\frac{1}{2}}]}{(1-\theta)^{(\tau-1)/2}}$$

$$= \frac{2(1-\theta)^{\tau/2}}{\alpha\theta \; K_\tau[\alpha(1-\theta)^{\frac{1}{2}}]} \quad [1 - K_{\tau-1}(\alpha)] \quad \frac{K_{\tau-1}[\alpha(1-\theta)^{\frac{1}{2}}]}{(1-\theta)^{(\tau-1)/2}}$$

$$= \frac{2(1-\theta)^{\frac{1}{2}}}{\alpha\theta} \quad [1 - K_{\tau-1}(\alpha)] \quad \frac{K_{\tau-1}[\alpha(1-\theta)^{\frac{1}{2}}]}{K_\tau[\alpha(1-\theta)^{\frac{1}{2}}]}$$

For shifted inverse Gaussian-Poisson, $\tau = -1/2$ and its $E[1/X]$ is

$$\frac{2(1-\theta)^{\frac{1}{2}}}{\alpha\theta} \quad [1 - K_{-3/2}(\alpha)] \quad \frac{K_{-3/2}[\alpha(1-\theta)^{\frac{1}{2}}]}{K_{-\frac{1}{2}}[\alpha(1-\theta)^{\frac{1}{2}}]}$$

But,

$$K_{-v}(z) = K_v(z) \; ;$$

$$K_{v+1}(z) = (2v/z) \; K_v(z) + K_{v-1}(z) \; ; \text{ and}$$

$$K_{\frac{1}{2}}(z) = (\pi/2z)^{\frac{1}{2}} \; e^{-z}$$

Hence,

$$E[1/X] = \frac{2(1-\theta)^{\frac{1}{2}}}{\alpha\theta} [1 - ((\alpha+1)/\alpha)(\pi/2\alpha)^{\frac{1}{2}} \; e^{-\alpha}] \quad \frac{\alpha(1-\theta)^{\frac{1}{2}}+1}{\alpha(1-\theta)^{\frac{1}{2}}}$$

$$= \frac{2\,[\alpha(1-\theta)^{\frac{1}{2}}+1]}{\alpha^2\,\theta}\,[1 - ((\alpha+1)/\alpha)(\pi/2\alpha)^{\frac{1}{2}}\,e^{-\alpha}]$$

(iii) Shifted Negative Binomial

$$P(X = j) = \binom{v+j-2}{j-1} p^v (1-p)^{j-1} \quad ; \; j = 1,2, \ldots$$

$$E[1/X] = \sum_{j=1}^{\infty} (1/j) \binom{v+j-2}{j-1} p^v (1-p)^{j-1}$$

$$= \sum_{j=1}^{\infty} (1/j) \frac{\Gamma(v+j-1)}{\Gamma(v)\,(j-1)!} p^v (1-p)^{j-1}$$

$$= \sum_{j=1}^{\infty} \frac{\Gamma(v+j-1)}{\Gamma(v)\,j!} p^v (1-p)^{j-1}$$

$$= \sum_{j=1}^{\infty} \frac{\Gamma[j+(v-1)]}{(v-1)\,\Gamma(v-1)\,j!} p^v (1-p)^{j-1}$$

$$= \frac{p}{(1-p)(v-1)} \sum_{j=1}^{\infty} \frac{\Gamma[j+(v-1)]}{\Gamma(v-1)\,j!} p^{v-1} (1-p)^j$$

$$= \frac{p}{(1-p)(v-1)} (1 - p^{v-1})$$

[Note : $\sum_{j=0}^{\infty} \frac{\Gamma[j+(v-1)]}{\Gamma(v-1)\,j!} p^{v-1} (1-p)^j = 1$]

REFERENCES

Allison, P.D. (1980). Inequality and scientific productivity. Social Studies of Science, 10, 163-179.

Balog, C. (1980). Multiple authorship and author collaboration in agricultural research publications. Journal of Research Communication Studies, 2(3), 159-168.

Bayer, A.E. (1982). A bibliometric analysis of marriage and family literature. Journal of Marriage and the Family, 44, 527-538.

Beaver, D. deB. & Rosen, R. (1978). Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. Scientometrics, 1(1), 65-84.

Beaver, D. deB. & Rosen, R. (1979). Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. Scientometrics, 1(2), 133-149.

Berenson, M.L., Levine, D.M. & Goldstein, M. (1983). Intermediate statistical methods and applications. Englewood Cliffs : Prentice-Hall, Inc.

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

Chatfield, C. & Goodhardt, G.J. (1970). The beta-binomial model for consumer purchasing behaviour. Applied Statistics, 19, 240-250.

Chen, Y. & Leimkuhler, F.F. (1986). A relationship between Lotka's law, Bradford's law, and Zipf's law. Journal of the American Society for Information Science, 37(5), 307-314.

Clarke, B.L. (1964). Multiple authorship trends in scientific papers. Science, 143(3608), 822-824.

Conover, W.J. (1980). Practical nonparametric statistics. 2nd edition. New York : John Wiley & Sons.

Consul, P.C. & Jain, G.C. (1973). A generalization of the Poisson distribution. Technometrics, 15(4), 791-799.

Consul, P.C. & Shoukri, M.M. (1984). Maximum likelihood estimation for the generalized Poisson distribution. Communications in Statistics - Theory and Methods, 13(12), 1533-1547.

Diamond, A.M. Jr. (1985). The money value of citations to single-authored and multiple-authored articles. Scientometrics, 8(5/6), 315-320.

Fox, M.F. & Faver, C.A. (1984). Independence and cooperation in research : The motivations and costs of collaboration. Journal of Higher Education, 55, 347-359.

Frame, J.D. & Carpenter, M.P. (1979). International research collaboration. Social Studies of Science, 9(4), 481-498.

Gladding, S.T. (1984). Multiple authorship in the Personnel and Guidance Journal : A 12-year study. Personnel and Guidance Journal, 62(16), 628-630.

Goffman, W. & Warren, K.S. (1980). Scientific information systems and the principle of selectivity. New York : Praegler Publishers.

Gordon, M.D. (1980). A critical reassessment of inferred relations between multiple authorship, scientific collaboration, the production of papers and their acceptance for publication. Scientometrics, 2(3), 193-201.

Haitun, S.D. (1982a). Stationary scientometric distributions : Part I. Different approximations. Scientometrics, 4(1), 5-25.

Haitun, S.D. (1982b). Stationary scientometric distributions : Part II. Non-Gaussian nature of scientific activities. Scientometrics, 4(2), 89-104.

Hefner, A.G. (1981). Funded research, multiple authorship and subauthorship collaboration in four disciplines. Scientometrics, 3(1), 5-12.

Hirsch, W. & Singleton, J.F. (1964). Research support, multiple authorship, and publications in sociological journals, 1936-1964. Unpublished paper, Purdue University.

Hodder, P. (1980). Limits to collaborative authorship in science publishing. Journal of Research Communication Studies, 2(3), 169-178.

Irwin, J.O. (1975). The generalized Waring distribution. Part II. Journal of the Royal Statistical Society A, 138, Part 2, 204-227.

Jain, G.C. & Consul, P.C. (1971). A generalized negative binomial distribution. **SIAM Journal of Applied Mathematics**, 21(4), 501-513.

Johnson, N.L. & Kotz, S. (1969). **Discrete Distributions**. New York : John Wiley & Sons.

Kendall, M.G. (1961). Natural law in the social sciences. **Journal of the Royal Statistical Society, Series A**, 124(Part 1), 1-16.

Kuch, T.D.C. (1978). Relation of title length to number of authors in journal articles. **Journal of the American Society for Information**, 29(4), 200-201.

Kumar, A. & Consul, P.C. (1979). Negative moments of a modified power series distribution and bias of the maximum likelihood estimator. **Communication in Statistics - Theory and Methodology**, A8(2), 151-166.

Lawani, S.M. (1980). **Quality, collaboration and citations in cancer research : A bibliometric study**. Ph.D. dissertation, Florida State University.

Lawani, S.M. (1986). Some bibliometric correlates of quality in scientific research. **Scientometrics**, 9(1/2), 13-25.

Lindsey, D. (1978). **The scientific publication system in social science**. San Francisco : Jossey-Bass.

Lindsey, D. (1980). Production and citation measures in the sociology of science : The problem of multiple authorship. **Social Studies of Science**, 10(2), 145-162.

Lodahl, J.B. & Gordon, G. (1972). The structure of scientific fields and the functioning of university graduate departments. **American Sociological Review**, 37(February), 57-72.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. **Journal of the Washington Academy of Science**, 16(12), 317-323.

Mandelbrot, B. (1953). An information theory of the statistical structure of language. In : Jackson, W. ed. **Proceedings of the Symposium on Applications of Communications Theory**. London : Butterworths; 486-500.

McCauley, J.E. (1963). Multiple authorship. **Science**. 141(3580), 579.

Meadows, A.J. (1974). **Communication in science**.

London : Butterworths.

Mendenhall, M. & Higbee, K.L. (1982). Recent trends in multiple authorship in psychology. Paper presented at the annual meeting of the Western Psychological Association, Sacramento, CA.

Mitchell, M.B. (1961). Trends towards multiple authorship in scientific publication. Journal of Psychology, 52, 125-131.

Mulkway, M.J. & Edge, D.O. (1973). Cognitive, technical and social factors in the growth of radio astronomy. Social Science Information, 12(6), 25-61.

Oromaner, M.J. (1975). Collaboration and impact : The career of multiple-authored publications. Social Science Information, 14(1), 147-155.

Over, R. & Smallman, S. (1970). Citation idiosyncracies. Nature, 228, 1357.

Over, R. (1982). Collaborative research and publication in psychology. American Psychologist, 37(9), 996-1001.

Pao, M.L. (1982). Collaboration in computational musicology. Journal of the American Society for Information Science, 33(1), 38-43.

Patel, N. (1973). Collaboration in the professional growth of American sociology. Social Science Information, 12(6), 77-92.

Pravdic, N. & Oluic-vukovic, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. Scientometrics, 10(5-6), 259-280.

Presser, S. (1980). Collaboration and the quality of research. Social Studies of Science, 10(1), 95-102.

Price, D. J. DE S. (1963). Little science big science. New York : Columbia University Press.

Price, D.J. DE S. & Beaver, D. deB. (1966). Collaboration in an invisible college. American Psychologist, 21, 1011-1018.

Schreiner, W., Kramer, M., Kriscer, S. & Langsam, Y. (1985). Nonlinear least-squares fitting. PC TECH JOURNAL, 3(5), 170-181.

Schweser, C. (1983). The economics of academic publishing. **Journal of Economic Education**, 14(1), 60-64.

Sichel, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In : Laubscher, N.F. ed. **Proceedings of the Third Symposium on Mathematical Statistics.** Pretoria : C.S.I.R.; 51-97.

Sichel, H.S. (1985). A bibliometric distribution which really works. **Journal of the American Society for Information Science**. 36(5), 314-321.

Simon, H.A. (1955). On a class of skew distribution functions. **Biometrika**, 42(3/4), 425-440.

Smart, J.C. & Bayer, A.E. (1986). Author collaboration and impact : A note on citation rates of single and multiple authored articles. **Scientometrics**, 10(5-6), 297-305.

Smith, S. (1958). The trend toward multiple authorship in psychology. **American Psychologist**, 13, 596-599.

Stefaniak, B. (1982). Individual and multiple authorship in chemistry and physics. **Scientometrics**, 4(4), 331-337.

Stein, G.Z., Zucchini, W. & Juritz, J.M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. **Journal of the American Statistical Association**, 82(Sept.), 938-944.

Subramanyam, K. & Stephens, E.M. (1982). Research collaboration and funding in biochemistry and chemical engineering. **International Forum on Information and Documentation**, 7(4), 26-29.

Subramanyam, K. (1983a). Bibliometric studies of research collaboration : A review. **Journal of Information Science**, 6(1), 33-36.

Subramanyam, K. (1983b). Collaborative publication and research in computer science. **IEEE Transactions on Engineering Management**, 30(4), 228-247.

Utz, W.R. (1981). Joint authorship. **Notices of the American Mathematical Society**, 28(5), 424.

Worthen, D.B. (1978). Short-lived technical literatures : A bibliometric analysis. **Methods of Information in Medicine**, 17(3), 190-198.

Zipf, G.K. (1949). **Human behavior and the principle of least** effort. New York : Addison-Wesley Press, Inc.

Zuckerman, H.A. (1967). Nobel laureates in science : Patterns of productivity, collaboration and authorship. **American Sociological Review**, 32, 391-402.

Zuckerman, H.A. (1968). Patterns of name ordering among authors of scientific papers : A study of social symbolism and its ambiguity. **American Journal of Sociology**, 74, 276-291.