INTRODUCTION TO INFORMETRICS

Quantitative Methods in Library, Documentation and Information Science

Leo EGGHE
Limburgs Universitair Centrum
Universitaire Campus
Diepenbeek, Belgium
and
Universitaire Instelling Antwerpen
Belgium

Ronald ROUSSEAU
Katholieke Industriële
Hogeschool West-Vlaanderen
Oostende, Belgium
and
Universitaire Instelling Antwerpen
Belgium

1990

ELSEVIER SCIENCE PUBLISHERS
AMSTERDAM • NEW YORK • OXFORD • TOKYO

ELSEVIER SCIENCE PUBLISHERS B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Distributors for the United States and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
655 Avenue of the Americas
New York, N.Y. 10010, U.S.A.

ISBN: 0 444 88493 9

PRINTED IN THE NETHERLANDS

PREFACE

In 1984, we two authors entered the field of bibliometrics - we now prefer the more generally accepted term 'informetrics' - when we became actively involved in the programme for higher education in library and documentation science organised by the University of Antwerp upon the initiative of the Flemish Interuniversity Council (VLIR). Before long, interest in a textbook on informetrics was evidenced not only in Flanders, but also in the Netherlands. Indeed, as a result of the EC-financed Erasmus project, Leo Egghe was charged with the responsibility for the informetrics programme at the University of Amsterdam as well.

Using one of Egghe's short Flemish courses as our starting point, we began working on the present volume in around 1986. Nowadays the field of informetrics has become so broad that no introductory book can aim at completeness. Still, we have tried to cover as many topics as possible. We occasionally resorted to the use of 'Notes and comments' to refer the reader to further developments which could not be covered fully in the main text.

In writing this book, we aimed at producing a clearly written text, with the topics presented in a logical format, a book which would appeal to the non-specialist (and the non-mathematician). In persuing this objective, we were confronted by a literature dispersed over a variety of very differently oriented journals and books. Moreover, as is well known, most scientists working in this field are informetricians only as a 'second choice'. They were educated as librarians, physicists, chemists, mathematicians, sociologists, psychologists or computer scientists, and their different backgrounds are revealed in their publications. Therefore, unifying these various points of view was not an easy task.

We expect this book to be of help to the informetrics teacher in organising his or her course and to be interesting and useful both as a course book and as background reading for students in library and information science. We hope in addition that practicing librarians will also find it useful, as we included many simple, non-mathematical library management techniques. Researchers and scholars working in the area of science policy may also find something of interest, since a lot of recent material has been included.

The book is divided into four parts, each containing a number of chapters, sections, subsections, and where necessary, sub-subsections. To refer to these subdivisions we use a decimal system, e.g. 'I.4.3.3' means Part I, Chapter 4, Section 3, Subsection 3. Equations, tables and figures are numbered in the same way, up to the section level. Consequently, '[I.5.18]'

means the 18th equation in Chapter I.5. Numbers for tables and figures are preceded by the words 'Table' or 'Fig.' (for definitions as well as references). Therefore, 'Fig.II.4.1' means the first figure in Chapter II.4. Table captions are written above the table and figure legends below the figure. References, which are listed alphabetically, are given in the following form : 'Name (year)'. For authors not appearing in the list of references as a senior author, there is a 'see' reference; for authors appearing as both a senior author and a secondary author (of a different paper), we added a 'see also' reference. While the reference list is long and covers all aspects of informetrics, it is, of course, not meant to be exhaustive. We apologise to any authors who have been unjustly omitted.

We are grateful to Prof. H.D.L. Vervliet, the founder of the programme for higher education in documentation and library science in Flanders. We are also greatly indebted to Prof. B.C. Brookes who encouraged us to write this book and who is really the great champion of the term 'informetrics'. Prof. Brookes was also Egghe's Ph.D. supervisor at the City University of London (UK). This thesis forms the basis for Part IV, dealing with informetric 'laws'.

Our sincere thanks are also extended to all those eminent scientists with whom we have had many lively contacts. We single out for mention here : A. Bookstein, T. Braun, Q.L. Burrell, W. Glänzel, D. Kraft, F.F. Leimkuhler, I.K. Ravichandra Rao, S.E. Robertson, G. Salton, J. Tague, R. Todorov and A.F.J. Van Raan.

Our appreciation is given to our institutes, the Limburgs Universitair Centrum, the Universitaire Instelling Antwerpen and the Katholieke Industriële Hogeschool West-Vlaanderen, for their interest and support. We also wish to thank the Belgian National Science Foundation, which has supported us on various occasions.

We thank the typist, Mrs. Reynders, for the excellent production of the camera-ready copy.

Lastly, it is a great pleasure to acknowledge the pleasant working relationship with Elsevier Science Publications, and in particular with Heleen van Gelderen and Susan Massotty.

The authors welcome any criticisms, corrections, additions or any other form of comments on the book in its present form. As one final word to our readers, we add the wish that the book will be useful to many persons, involved in all kinds of information work.

Leo Egghe
Diepenbeek, Belgium

Ronald Rousseau
Stene, Belgium

December 1989

TABLE OF CONTENTS

0. INTRODUCTION

There is no measurement (i.e. meaningful data) without theory and no theory without data. Although this statement may appear to be a vicious circle, it is not. What we mean is that there is an ongoing infinite spiral, in which more and more refined theories are being tested better and better through more refined measurements.

For the communication of scientific insights, experience and discussions, logic and mathematics are indispensible. Mathematical aids allow models to be constructed and measurements to be made. On the other hand, meaningful measurements are only possible because certain laws and theories, whether they are deterministic or probabilistic, exist. We can understand our measurements only because we understand theory (at least partially). Thanks to the advent of the computer, it has now become easier to collect data in libraries or from sources stored on computer, so that models are being constructed in what we for the time being will refer to as 'Library and Information Science'.

The scientific method of the theory-measurement cycle is very powerful, but we have to pay a price for it. As we define our theoretical models more precisely, they become detached from the real world. This is why we have to supplement our mathematical models, definitions and theories with verbal interpretations, again using concepts that can be understood intuitively, but which are slightly ambiguous and inaccurate.

In the field of 'Library and Information Science' model-building has only just begun. Very often we have to be satisfied with elementary data collection and an intuitive explanation. Yet already, we are hearing complaints from the uninitiated that theories are already too abstract and not really applicable. Perhaps a book like this can be of help, before theory really takes off. Indeed, characteristic of the immature state of our science, even its very name is still a subject of debate. Should one use the term 'bibliometrics' or "scientometrics' or 'informetrics' (perhaps even 'librametrics')? And what does the term cover? Does it include science policy issues, theoretical aspects of information retrieval, some artificial intelligence techniques, theories of questionnaires?

In our view, informetrics deals with the measurement, hence also the mathematical theory and modelling of all aspects of information and the storage and retrieval of information. It is mathematical meta-information, i.e. a theory of information on information, scientifically developed with the aid of mathematical tools (cf. Burton (1988)). See, for example, Nacke (1979) and Bonitz (1982) for an early mention of the term 'informetrics'.

Historically, bibliometrics developed mainly in the West, and arose from statistical studies of bibliographies. Before the term 'bibliometrics' was proposed by Pritchard (1969), the term 'statistical bibliography' was in some use. According to Pritchard (1969), it was Hulme (1923) who initiated the term 'statistical bibliography'. Hulme used the term to describe the process of illuminating the history of science and technology by counting documents.

Pritchard's timely proposal caught on immediately, but the content of the term remained somewhat of a problem (Broadus, 1987). According to Pritchard, bibliometrics means the application of mathematics and statistical methods to books and other communication media.

On the other hand, the term 'scientometrics' (derived from the Russian 'naukometria') was used mainly in the East and is defined as the study of the measurement of scientific and technological progress. This also explains the foundation in 1978 and the title of the journal 'Scientometrics' in Hungary. For more information on the history and the contents of these names we refer the reader to Egghe (1988f).

We fully agree with Brookes (1988b) that the term 'bibliometrics' ties us too narrowly to libraries and the documentary origin of the field. Hence, we will restrict this term to the mathematical study of libraries and bibliographies. Scientometrics, on the other hand, deals mainly with science policy applications. Therefore, we support Brookes (1988b) who advocates the use of the term 'informetrics', a term which takes cognizance of the fact that modern technology has imposed on us new non-documentary forms of knowledge representation and of its transmission and dissemination. Scientists such as Dou (Dou et al. (1988)) define even the term 'bibliometrics' as the statistical treatment of downloaded data. Although we do not agree with this definition, it is symptomatic of the influence computer technology exerts on our field.

We harbour no strong feelings about the vagueness of the term. Do chemists and physicists quarrel about chemical physics? Is it important to determine whether a certain paper should be considered as a mathematical paper, an econometric one, or even an economics paper? Every new field has vague boundaries and even established fields such as physics and chemistry cannot be separated in a clear way. So, is not informetrics simply that which informetricians do?

Of course it is certainly important to clearly define the main problems in the field (Brookes (1988a)), to look for new, important applications (Tague (1988)), to pay more attention to the modelling process (Leimkuhler (1988)) and to make use of dedicated software in computer experiments.

In an attempt to give informetrics a place among other fields we present the following diagram (Fig. 0.1) :

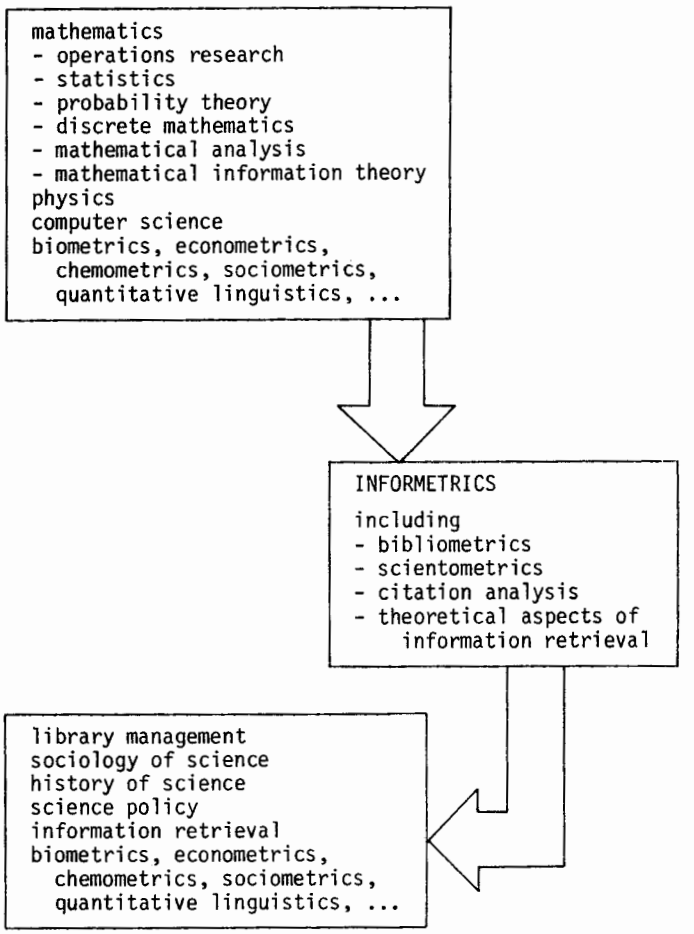

Fig. 0.1

In this diagram, we mean that informetrics borrows tools (techniques, models, analogies) from mathematics, physics, computer science and the other -metrics. On the other hand, informetrics is used in or applied to : library management, the sociology of science, the history of science, science policy and information retrieval. Moreover, we feel that a real interaction between informetrics and biometrics, econometrics, chemometrics, quantitative linguistics and so on would be very beneficial for all fields involved. Until now there has only been

a (small) influx from the other fields into informetrics, but we are certain that our field also has something important to offer to others.

Finally, a short summary of the book. Part I covers statistical methods. It starts with elementary descriptive statistics and elements of probability. It continues with an important chapter on inferential statistics (hypothesis testing), including regression, correlation and nonparametric statistics. Next, there is a chapter on sampling theory, including overlap problems. Part I concludes with several techniques of multivariate statistics : multivariate regression and correlation, principal component analysis, multidimensional scaling and cluster techniques.

Part II deals with operations research and library management. Applications of linear programming (including transportation and assignment problems) are given, followed by elements of queueing theory. Special attention is paid to book circulation interference.

Part III handles citation analysis : citers' motivations, citation networks, bibliographic coupling and co-citation analysis, JCR and its citation measures and obsolescence. Some aspects of science policy applications are also studied.

Finally, part IV deals with informetric models and their interrelationships. At the heart of this theory is the dual approach between sources and items giving rise to the definition of 'Information Production Processes'. Explanations and applications of the classical informetric laws as well as fitting methods are provided.

I. STATISTICS

I.0. INTRODUCTION

Statistics play a vital role in the development of informetrics as an academic discipline and, more importantly, as a practical discipline. The aim of this part is to introduce the reader to some of the concepts and methods used in statistical analysis.

Library automation provides managers with an increasing amount of data. If the head of a library or documentation centre wishes to convert this mass of numbers into useful information he or she needs ways to summarise large sets of data. Descriptive statistics (Chapter I.1) can help to fulfil this aim.

One of the fundamental concepts of statistics is probability (Chapter I.2). All statistical tests involve the calculation of probabilities, either directly or indirectly. Statistical hypotheses are never said to be true or false. Instead, the probability that they are true or false is stated. We will outline some simple rules of probability and introduce a number of theoretical probability distributions.

A central aspect in discovering new knowledge about the real world consists of observing some arbitrary elements of the set of objects, events or persons under discussion : a so-called random sample (Chapter I.4). On the basis of this sample, one makes a statement about the totality of elements (the population). This part of statistics is called 'inferential statistics' (Chapter I.3 and I.5).

The primary topics in inferential statistics are the testing of hypotheses, regression, goodness-of-fit tests, the analysis of contingency tables and multivariate techniques such as principal components analysis, multidimensional scaling and hierarchical cluster analysis.

The examples given in the text have been kept deliberatedly simple, and when no special mention is made of how and where data were collected, this means that they have been conceived for illustrative purposes. The reader will, however, find numerous references to the literature on informetrics.

An excellent critical review of statistical methods in information science research can be found in Kinnucan et al. (1987).

I.1. DESCRIPTIVE STATISTICS

The term 'descriptive statistics' refers to a set of methods, procedures and techniques used to represent, summarise, or otherwise communicate the essential characteristics of a set of raw data. Some important aspects of descriptive statistics are tabular and graphical representations and the calculation of a single number representing a particular characteristic of the data in question. Applying the techniques of descriptive statistics allows one to make statistical inferences, e.g. the use of chance models to draw conclusions from data. These conclusions help library managers to solve the problems they are confronted with.

I.1.1. Tables

Questionnaires, tally sheets and computer printouts all generate data, usually numbers, in some form or another. Writers, whether they are penning scientific articles or popular journalism, often represent numerical data in tabular form. A table not only occupies less space than the narrative form, but it also enables figures to be located more readily and facilitates comparisons between different figures or sets of figures. To do this effectively, a table must be compiled with its future use in mind.

Although the numerical values in a data display are commonly referred to as 'the data', the numbers are only one element of the data. Indeed, all data refer to some real-world event; they also include the content elements, i.e. words and phrases that connect these numbers with the observed phenomenon. At the basic level of description, content elements are the familiar who, what, how, where and when - defined in more formal terms as observer, matter, function, space and time - and at the next level of description they relate to the aspect of reality measured and the set of reported values, Clark (1987).

Let us begin by taking a look at the well-known 'Applied Geophysics' data in Bradford's paper (1934) (Table I.1.1). The title of the table identifies the time period covered : 1928-1931, incl. The legend indicates the function, the matter and the observer : the production of papers (references) in journals.

The aspect of paper production being measured is Applied Geophysics : the table provides information on the number of papers in Applied Geophysics and their distribution over various journals. The data collector and place where the collecting occured are mentioned in the body of the paper : it was Mr. E. Lancaster Jones who carried out the investigation in the Science Museum Library.

Table I.1.2 (taken from Clark (1987), but slightly amended to include the place of publication) lists the categories of information needed to provide the

Table I.1.1. Bradford's Applied Geophysics data (with original legend)

Applied geophysics, 1928-1931, incl.				
A.	B.	C.	D.	E.
1	93	1	93	0
1	86	2	179	0.301
1	56	3	235	0.477
1	48	4	283	0.602
1	46	5	329	0.699
1	35	6	364	0.778
1	28	7	392	0.845
1	20	8	412	0.903
1	17	9	429	0.954
4	16	13	493	1.114
1	15	14	508	1.146
5	14	19	578	1.279
1	12	20	590	1.301
2	11	22	612	1.342
5	10	27	662	1.431
3	9	30	689	1.477
8	8	38	753	1.580
7	7	45	802	1.653
11	6	56	868	1.748
12	5	68	928	1.833
17	4	85	996	1.929
23	3	108	1065	2.033
49	2	157	1163	2.196
169	1	326	1332	2.513

Column A gives the number of journals producing a corresponding given number of references. Column B gives the corresponding number of references during the period surveyed. Column C gives the running sum of the numbers of Column A. Column D gives the running sum of the numbers of Column B multiplied by A. Column E gives the common logarithms of Column C numbers.

reader with a complete picture of the data, while Table I.1.3 shows the resulting descriptor set for Bradford's paper. Note the entry for 'aspect'. The term 'aspect' is, by definition, a relative term (it is always an aspect of something else), and the arrow points to its antecedent, the topic term in the descriptor set. Thus the entire entry, including the arrow, specifies not only what was measured, but why : the underlying question the data are designed to answer.

Ideally, tables should contain all the elements necessary to fill in all the entries in Nancy Clark's editorial table tamer (except, of cours, the place of publication when original data are submitted for publication).

Table I.1.2. Nancy Clark's 'editorial table tamer'

Current source	: Author of this representation of data; publication date and place
Source of data	: Data collector; time of data collection
Observer	: Respondent group, source of reported values
Matter	: Entity(ies) involved in the event discussed in the table
Function	: Nature of the event discussed
Space	: Location of this event
Time	: Time of this event
Aspect	: Aspect of reality + pointer to topic term
Domain	: Nature of values

Table I.1.3. Descriptor set for Bradford's Applied Geophysics data, composed by R. Rousseau

Current source	: S.C. Bradford; Engineering, 137 (1934) 85-86
Source of data	: E. Lancaster Jones; 1932 (?)
Observer	: journals
Matter	: papers
Function	: production (publication)
Space	: all over the world, but confined to those primary and abstracting journals available at the Science Museum Library
Time	: 1928-1931 incl. *
Aspect	: publications on the subject of Applied Geophysics [→ highly skewed distribution, described by a formula which later became known as Bradford's law]
Domain	: 1,93

* From the remainder of Bradford's paper we learn that the actual data-collection period included part of 1932.

Let us next take a look at tables from another perspective, namely that of the reader. Suppose a table of data is presented to you : how should you read this table in order to obtain as much information as possible, as quickly as possible? For this A. Ehrenberg (1986) presents the following general guidelines :

1. Take in the broad subject matter of the table and the variables, without yet worrying over details, sources, etc. (but if there is a text or commentary, one should probably glance at this first; it may provide access to what the table is saying).
2. Focus first on one row and/or one column, preferably of averages. Establish the range of variation, i.e. the highest and lowest readings, as mental markers.
3. Round all figures to one or two effective digits in one's head, to facilitate mental arithmetic and make the results more memorable.
4. Compare the detailed readings in the body of the table against these patterns as norms.
5. Consider the wider meaning of the results and do a more formal analysis.

I.1.2. Scales of measurement

In this section we briefly introduce the notion of scale. A more complete description can be found, for example, in Roberts (1979).

A *nominal scale* is used if observations are merely labelled (by a number or a name). The actual label has no significance (except possibly as a mnemonic aid), and any change of label will contain the same information. For instance, in scientometric investigations of countries, the names of the countries under study could constitute a nominal scale.

Opinions are frequently measured in terms of ordinal data. For instance, a library patron may be asked to rank the quality of various library services. The answer only shows the relative position of these services and not the extent to which one service is better than another. The *ordinal scale* provides information about the ordering of categories, but does not indicate the extent of the differences between observations.

An *interval scale* of measurements reveals more than the ordering of categories : it gives the difference between them with respect to a fixed but arbitrary origin and a fixed but arbitrary unit. A typical example of an interval scale is the common method of measuring temperature. When comparing Fahrenheit and Celsius degrees, we vary the origin and the units.

A *difference scale* gives the exact difference between certain categories; only the origin is arbitrary. A typical example is the calendar : a year is a meaningful unit but the fact that this is the year 1990 is purely arbitrary.

In *ratio scales*, the origin is fixed but the units are arbitrary. Mass defines a ratio scale, as it is possible to determine a zero point and then change the unit of mass by multiplying it by a positive constant. Temperature defines a ratio scale only if the Kelvin scale is used. The term 'ratio scale'

has been employed because ratios of quantities make sense on such a scale. We further note that the use of logarithms changes a ratio scale into a difference scale.

Finally, when the origin and the units are fixed, we have an *absolute scale*. Counting is done on an absolute scale.

I.1.3. Graphical representations

I.1.3.1. Frequency distributions, histograms and frequency polygons

For data not measured on a nominal or an ordinal scale, determining the *frequency distribution* is a means of imposing a certain structure on raw data. As an example, we consider the average number of references of source publications of the Science Citation Index (taken from Nakamoto (1988) ; further information about the Science Citation Index will be given in Section III.1.3).

Table I.1.4. Average number of references of source publications in the Science Citation Index

A. Year
B. Average number of references

A	B	A	B
1961	12.1	1973	12.3
1962	12.0	1974	13.1
1963	12.1	1975	13.2
1964	11.8	1976	13.7
1965	12.4	1977	14.9
1966	11.2	1978	15.2
1967	11.1	1979	15.0
1968	12.0	1980	15.9
1969	11.3	1981	16.1
1970	11.4	1982	15.5
1971	12.0	1983	15.4
1972	12.3	1984	15.7

For the type of investigation aimed at here, the year does not matter (column A); only the data in column B will be manipulated. To obtain a frequency distribution, we first group the data into convenient class intervals. Table I.1.5 lists the distribution of the data in Table I.1.4. To solve the problem of numbers falling on the boundary between two class intervals, we follow the convention of including the left end point in the class interval, but not the right end point.

Table I.1.5. Distribution of the data in column B from Table I.1.4

[11.0,12.1[	𝍸 |||	8
[12.1,13.2[	𝍸 |	6
[13.2,14.3[	||	2
[14.3,15.4[	|||	3
[15.4,16.5[	𝍸	5

We will next explain how to draw a *histogram* from this distribution table. The first step is to establish the horizontal axis. While it is convenient to have intervals with the same width (as in this example), many distributions encountered in real informetric situations (such as in the next example) do not lend themselves to such an approach. The simplified case depicted in Table I.1.5 can easily be drawn as the histogram in Figure I.1.1.

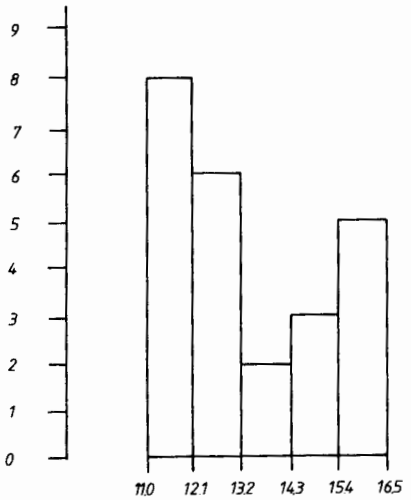

Fig.I.1.1. Histogram of the frequency distribution in Table I.1.5

Sometimes one may also wish to construct a *frequency polygon*, in which case, the same class intervals as in the histogram are used. The next step is to join the midpoints of the upper horizontal sides of the bars in the histogram. The two ends of the resulting polygon are then usually linked to the midpoints of the class intervals adjacent to the intervals used for the

histogram. When class intervals with the same width are used, it is easy to see that the area under the polygon is equal to the area under the corresponding histogram. As two or more frequency polygons can be displayed on the same graph - which is hardly possible for histograms - they are often preferred for purposes of graphic comparisons. Figure I.1.2 illustrates the frequency polygon for Table I.1.5.

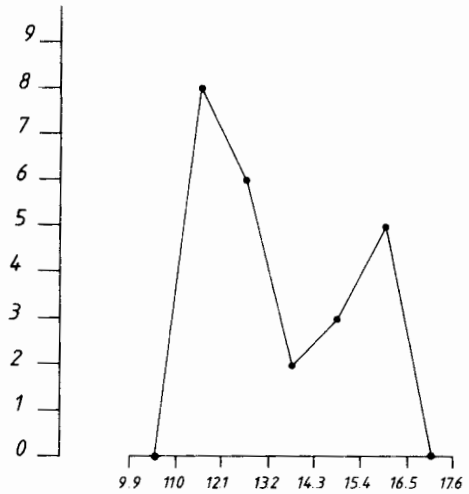

Fig.I.1.2. Frequency polygon of the distribution in Table I.1.5

As a second example we consider the number of citations in 1984 of papers published by the first 100 journals in the alphabetical list of the Journal Citation Reports, to papers published in the preceding year 1983. (For further information on the Journal Citation Reports, the reader is referred to Chapter III.5).

Table I.1.6 1984 citations of papers published by the first 100 journals in the alphabetical list of the Journal Citation Reports; covering papers published in 1983

54	29	4	9	9	3	33	2	32	68
161	446	14	7	228	0	18	10	5	43
2	14	1	129	189	13	18	4	4	6
3	1	0	138	443	3	36	8	5	6
2	2	7	63	93	2	2	13	71	52
384	111	2	13	129	6	6	33	85	130
23	91	7	7	14	75	15	5	147	10
22	5	2	2	0	58	0	102	154	15
13	5	0	87	448	24	131	225	67	52
23	11	28	1	8	20	339	86	67	48

Table I.1.7 Frequency table : distribution of citation data from Table I.1.6

[0,10[	40	[100,200[	11
[10,20[	14	[200,300[	2
[20,50[	13	[300,400[	2
[50,100[	15	[400,500[	3

We begin by drawing the x-axis as in Fig.I.1.3.

Fig.I.1.3 x-axis for the histogram of Table I.1.7

According to convention, a histogram represents percents by area. As a histogram consists of blocks, no special problems arise when all intervals are equal, as in the preceding example. In this case, however, we have class intervals of unequal lengths. So we have to adapt the height of each block in such a way that the area of each block represents the percentage of cases in the corresponding class interval. Applied to the citation data, this yields Fig.I.1.4. We note that drawing a vertical scale here would be nonsensical.

Lastly, for data on a nominal scale, classes naturally coincide with the labels of observations. Properly speaking, in this case we have *bar charts*, rather than histograms. Table I.1.8 depicts data for the daily number of book loans in two libraries.

Table I.1.8 Book loans in libraries A and B

Days	M	Tu	W	Th	F	Sa	Su
A	120	133	124	107	129	0	0
B	90	91	83	87	86	88	88

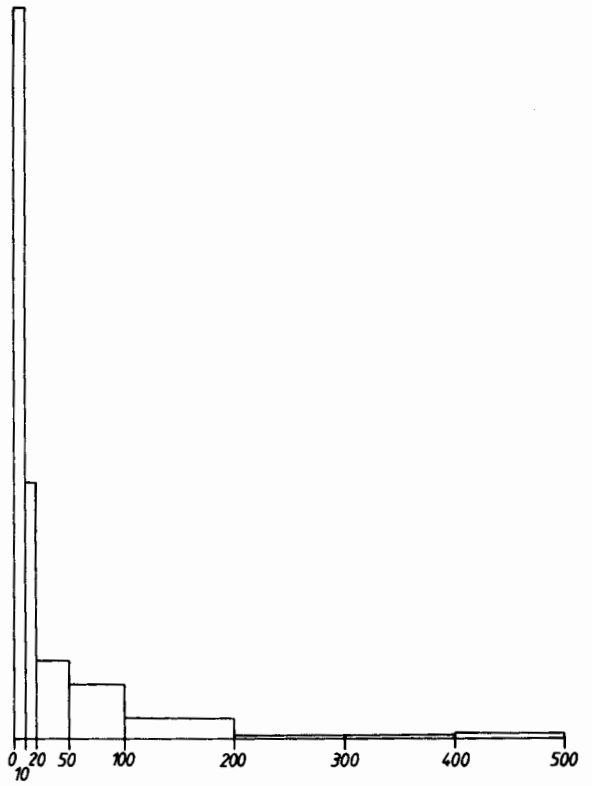

Fig.I.1.4 Histogram of citation data (Table I.1.7)

Fig.I.1.5 consists of histograms for libraries A and B (considered as nominal data). Indeed, Table I.1.8 can be viewed as a tabular representation

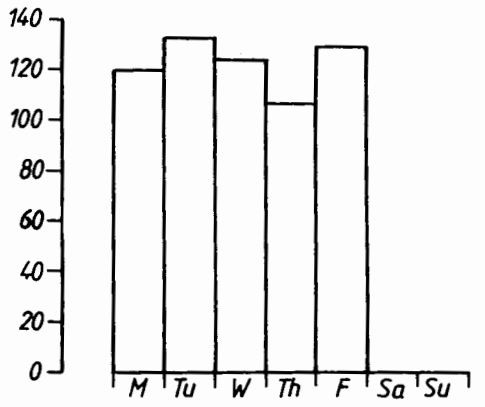

Fig.I.1.5.a Bar chart for book loans from library A

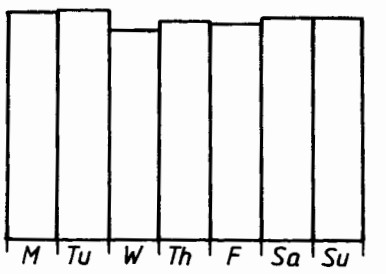

Fig.I.1.5.b Bar chart for book loans from library B

of two different kinds of measurement. In the first interpretation, the data are seen as days of the week. In this case, the total number of observations in both libraries is 613, and we have nominal data (when the natural order of days is ignored). This case is depicted in Fig.I.1.5. In the second interpretation, the data are viewed in terms of the numbers of book loans a day. In this case, the total number of observations is 7, and we are working on an absolute scale.

I.1.3.2. Logarithms and logarithmic representations

Many important informetric representations make use of graphs with a *logarithmic scale* on one or both axes.

I.1.3.2.1. Semi-logarithmic representations

In the case of *semi-logarithmic* representations, only one of the two axes is scaled in terms of the logarithm of the variable. Just as normal graphs can be drawn on special graph paper, semi-logarithmic graphs can be plotted on paper in which points (x,y) are actually represented by $(\log_{10} x, y)$ or $(x, \log_{10} y)$.

I.1.3.2.2. Logarithmic (also called double-logarithmic) representations

In the case of *double-logarithmic* representations, both axes are logarithmically scaled. On double-logarithmic paper, a point (x,y) is represented by $(\log_{10} x, \log_{10} y)$.

I.1.3.2.3. Practical use of logarithms

In general, the logarithmic method of plotting is used when relative changes are important, since equal linear displacements on a logarithmic scale indicate equal proportional changes in the variables themselves.

In informetric practice logarithms are mostly used to represent non-linear relations in a linear way. Consider, for example, the relation

$$y = C\, a^x \quad , \qquad [I.1.1]$$

where C and a are strictly positive constants. Taking logarithms of both sides results in :

$$\log_{10} y = \log_{10} C + x \log_{10} a \quad . \qquad [I.1.2]$$

Here we see that plotting (x, $\log_{10} y$) yields a straight line. In this case, it would be better to use semi-logarithmic paper with a logarithmic scale on the y-axis.

Let $y = D \log_a x + E$, in which a, D and E are constants (where $a > 0$). By using the general relation $\log_a x = \log_a b . \log_b x$ and taking b to be 10, we obtain the relation :

$$y = (D \log_a 10) \log_{10} x + E . \qquad [I.1.3]$$

This turns out to be a straight line when semi-logarithmic paper with a logarithmic scale on the x-axis is used.

Lastly, if $y = B x^a$, in which a and B are constants (where $B > 0$), taking logarithms yields :

$$\log_{10} x = \log_{10} B + a \log_{10} x . \qquad [I.1.4]$$

In this case, using double logarithmic paper results in a straight line.

I.1.3.3. Graphical representations : further remarks

Graphs are vital for communication in science : at best they can summarise vast amounts of quantitative information. Although graphic design as a means of communicating statistical information was established in the 1800's - with William Playfair (1786) being the most interesting precursor - the recent explosion in computer graphics software has led to an increasing use of graphs and made it easier to design new types.

A survey by Cleveland (1984a) showed that a significant number of graphs in scientific publications contain mistakes of some kind. Indeed, a detailed analysis of all the graphs in one volume of Science revealed that 30 % contained errors. This result was confirmed by Howarth and Turner (1987), who found that between 18 % and 35 % of the graphs in Geochemical journals contained at least one error. In both investigations errors were classified according to the following four types :

(a) Construction : a mistake in the construction of the graph, such as tick marks spaced incorrectly, mislabels, missing items and wrong scales.
(b) Degraded image : some aspect of the graph is missing or partially missing because of poor reproduction.
(c) Explanation : something on the graph is not explained.
(d) Discrimination : items on the graph, such as different symbol types, cannot be easily distinguished because of the design or size of the graph.

As a result of this survey Cleveland (1984a) concluded that graphical communication in science is badly in need of improvement. He pointed out five areas in which further research and development could assist this improvement, namely :

1) carrying out studies of how graphs are used;
2) developing new methods for data presentation;
3) developing guidelines;
4) studying human graphical perception;
5) developing software for statistical graphics.

Of these five areas the study of graphical perception is considered of fundamental importance. When a graph is made, various means are used to encode information on this graph, such as the positions of symbols, the lengths and slopes of line segments, areas and colour. A reader later studying this graph visually decodes the encoded information. This is what Cleveland and McGill (1987) call graphical perception. Studies in graphical perception should provide a scientific foundation for the construction of better statistical graphs.

We conclude this short survey on graphical representation by presenting a few guidelines (taken from Cleveland (1984a), Cleveland (1985) and Howarth and Turner (1987)) on how to make more effective graphs.

Guidelines :

1. When feasible, put important conclusions into graphical form. Most people do not read an entire article from beginning to end; readers skimming a paper are drawn toward graphs. Try to make graphs and their legends tell the story of your article.

2. Describe the graphs clearly. The combined information in the figure legend and the text in the body of the paper should provide a clear and complete description of everything on the graph. Detailed figure legends can often be of great help to the reader. First describe completely what is graphed in the display, then draw the reader's attention to salient features, and then briefly state the importance of these features.

3. Make the quantitative information that is graphed stand out. Be sure that different items on a graph can be easily visually distinguished.

4. Avoid cluttering graphical displays. For example, too much writing on the plotting region can interfere with viewers' perception of geometric patterns.

5. Subject to scaling constraints, choose the scales so that the data fill up as much of the data region as possible. Do not insist that zero always be included on a scale showing magnitude. Use a logarithmic scale when it is

important to understand percentage change or multiplicative factors; a logarithmic scale can moreover improve resolution. Use a scale break only when necessary, and if a break cannot be avoided use a full scale break.

6. Make graphs visually clear and capable of withstanding reduction. For example, lines must be thick enough, letters must be large enough, and plotting symbols must be large enough to accomodate the reduction.

7. Finally, recall that graphing data is an iterative, experimental process and proofread graphs just as any other part of a published manuscript.

For further information on this topic we refer the reader to Cleveland (1979), Cleveland, Harris and McGill (1983), Tufte (1983), Cleveland (1984b), Cleveland and McGill (1984), Cleveland and McGill (1986), Clark (1987) and Becker and Cleveland (1987). We especially recommend Cleveland's book 'The Elements of Graphing Data' to anyone seriously interested in improving the clarity of graphical representations of scientific data.

I.1.4. Measures of central tendency

The formulas in this section apply to a sample or to a population of size N, denoted x_i, i = 1,2,...,N. How to draw such a sample will be discussed in Chapter I.4.

I.1.4.1. The *mean* (also called the *average* or the *arithmetic mean*). For data not measured on a nominal or an ordinal scale the number

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad [I.1.5]$$

is called the mean of the sample (or the population). We will often use μ to denote the mean. Tables I.1.4B, I.1.6, I.1.8A and I.1.8B (second interpretation) contain respectively the following means : 13.2, 58.4, 87.6 and 87.6.

I.1.4.2. Weighted mean

If a weighting factor $w_i \geq 0$ is associated with every value x_i then $W = \sum_{i=1}^{N} w_i$ is the total weight, and

$$\bar{x} = \frac{1}{W} \sum_{i=1}^{N} w_i x_i \qquad [I.1.6]$$

is the *weighted mean*. The (unweighted) mean of I.1.4.1 can be considered as a weighted mean, where every $w_i = 1$.

I.1.4.3. Geometric mean

If all x_i are strictly positive, then the *geometric mean* (GM) is defined as

$$GM = \sqrt[N]{x_1 . x_2 . \ldots . x_N} \quad , \qquad [I.1.7]$$

i.e. the N-th root of the product of the N values. To compute GM, it is often easier first to compute the mean of the logarithms of the x_i and then to take its antilogarithm :

$$GM = 10^m \quad ,$$

where $m = \frac{1}{N} \sum_{i=1}^{N} \log_{10} (x_i)$.

The geometric mean is useful in averaging ratios, percentages and rates. The geometric mean of Table I.1.4B is 13.13.

I.1.4.4. Harmonic mean

If all x_i are strictly positive, then the *harmonic mean* (HM) is defined as

$$HM = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \quad . \qquad [I.1.8]$$

An example of the use of the harmonic mean is given by Zusne (1976), who used the harmonic mean of the ages of the author's first and last publications to predict the peak of creativity of outstanding psychologists. For Table I.1.4B, the HM is 13.03.

I.1.4.5. Arithmetic, geometric and harmonic mean are related through the following inequality :

$$HM \leq GM \leq \bar{x} \quad ,$$

where the equality signs hold if and only if all values are equal.

I.1.4.6. Median

If the data are arranged in descending order of magnitude, then the *median* Md is given by the (N+1)/2-nd value. When N is even, the median is usually

taken as the mean of the two middle values of the set of ordered data. A median divides the area under the frequency polygon into two equal parts. The medians of Tables I.1.4B and I.1.6 are respectively 12.35 and 14.5.

I.1.4.7. Mode

A *mode* Mo of a sample of size N is the most frequently occurring value, i.e. the most common value. A mode may not exist at all (e.g. when all observations are different) and even if it does exist, it may not be unique. For nominal scales, the mode is the only meaningful measure of central tendency. The mode of Table I.1.6 is 2. However, in this case it makes more sense to use the term 'modal class'. The modal class is then the first class : [0,10[. Similarly, for Table I.1.4B, the modal class is the first one : [11.0,12.1[. Finally, for both libraries (Table I.1.8, first interpretation), the mode is Tuesday.

I.1.4.8. Applications

Otherwise irregular data curves can be made more regular by using averages. They make the general trend more conspicious. As such, the use of averages can be considered as a smoothing technique. A good example of this is given in Baglow and Bottle (1979) : see Fig.I.1.6.

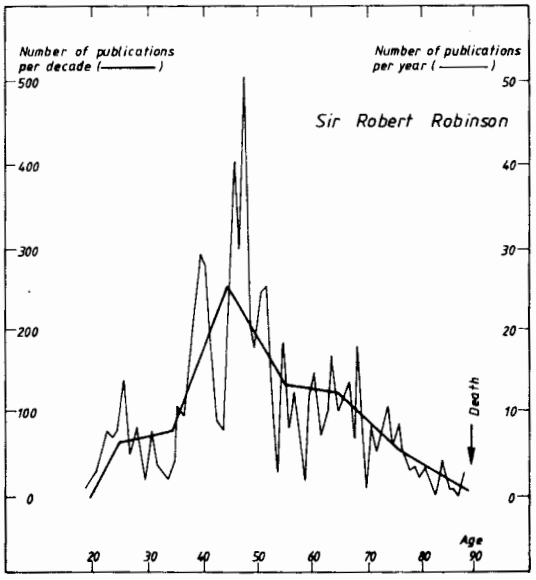

Fig.I.1.6 Use of averages as a smoothing technique : publications by Sir Robert Robinson

A more elaborate use of weighted averages as a smoothing technique is described by Winston (1984; p.77) and has been applied to citation data in Rousseau (1989a).

To apply this method, data $(x_i)_{i=1,\ldots,N}$ are ordered in a natural way : the index i denotes, say, time or locations on a line. Moreover, the x_i are only known up to a limited certainty, indicated by a confidence index c_i, $0 \leq c_i \leq 1$. Then the following relaxation procedure is used :

$$x_i^{(k)} = c_i x_i^{(0)} + (1-c_i)\,\frac{1}{2}\,(x_{i+}^{(k-1)} + x_{i-}^{(k-1)}) \ .$$

Here $x_i^{(k)}$ denotes the smoothed value of x_i after k iterations; $x_i^{(0)}$ is the starting value of x_i, while x_{i+} is the right neighbouring value of x_i and x_{i-} is the left neighbouring value. For end points (i=1 and i=N) the formula is changed into :

$$x_1^{(k)} = c_1 x_1^{(0)} + (1-c_1)\, x_{1+}^{(k-1)}$$

and

$$x_N^{(k)} = c_N x_N^{(0)} + (1-c_N)\, x_{N-}^{(k-1)} \ .$$

Note that these equations consist of two terms. The first is the measured value, multiplied by its confidence index. This part does not change during the iteration procedure. The higher the confidence index of this value is, the more important the term is. The second term is determined by the actual value of neighbouring data. At any moment a new iterated value can be obtained as the weighted average of the old value and the old values of neighbouring points. This procedure usually converges quickly to a stable state.

I.1.5. Measures of dispersion

Measures of central tendency such as the mean are not sufficient to describe data. A good example is given by Table I.1.8. Here $\bar{x}$ does not even give a clear impression of how many books are loaned out every day : although the average number of book loans is the same for both libraries, the lending pattern is completely different (if for no other reason than the fact that library A is closed during the weekend - let us say A is a business library and library B is not). We observe that for nominal or ordinal data the notion of dispersion makes no sense.

I.1.5.1. Variance and standard deviation

The most commonly used measures of dispersion are the *variance* (denoted as σ^2) and its square root, the *standard deviation* (σ).

The variance of a set of data (x_i), where i = 1,...,N, is defined as

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \ . \qquad [I.1.9]$$

The variance can also be expressed in the following ways :

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 \qquad [I.1.10]$$

as well as

$$\sigma^2 = \frac{1}{2N^2} \sum_{k=1}^{N} \sum_{i=1}^{N} (x_k - x_i)^2 \ . \qquad [I.1.11]$$

The standard deviation σ is nothing but the square root of the variance. Mean, median and standard deviation are related through the equation :

$$|\mu - Md| \leq \sigma \ .$$

For a proof the reader is referred to, for example, Falk (1981).

For the data in Tables I.1.4B, I.1.6, I.1.8A and I.1.8B the variances are respectively 2.86, 9311.5, 3124.8 and 5.96; hence the standard deviations are : 1.69, 96.5, 55.9 and 2.44.

I.1.5.2. Range

This is the simplest measure of dispersion. It is defined as the difference between the highest and the lowest value of a variable observed in an experiment. Although the range is easy to compute, it depends only on two extreme values and does not take the distribution into account. This means that it can be heavily influenced by sampling fluctuations, so that the range is only a crude measure of dispersion.

For the data in Tables I.1.4B, I.1.6, I.1.8A and I.1.8B the ranges are respectively 5.0, 448, 133 and 8.

I.1.5.3. Mean deviation

The *mean deviation* is defined as :

$$MD = \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}| \quad .$$

This measure is not often used in practice.

I.1.5.4. Interquartile range

When data are arranged in order of magnitude, the j^{th} *quartile* Q_j, where $j = 1,2,3$, is given by the $j(N+1)/4$-th value. Again, it may be necessary to interpolate between successive values. The second quartile is the median.

Similarly, the j^{th} *percentile* P_j, where $j = 1,\ldots,99$, is given by the $j(N+1)/100$-th value. Note that $P_{25} = Q_1$, $P_{50} = Q_2 = Md$ and $P_{75} = Q_3$.

The interquartile range is $Q_3 - Q_1$ or $P_{75} - P_{25}$ and may be considered as a refinement of the range.

I.1.5.5. Coefficient of variation

We define the *coefficient of variation* V as $\frac{\sigma}{\mu}$. This measure of dispersion will play an important role in the study of inequality in informetrics (an aspect of informetrics closely related to econometrics). For this, see Part IV.

I.1.5.6. Moments

a) The r^{th} *moment about the origin* is given by

$$m'_r = \frac{1}{N} \sum_{i=1}^{N} x_i^r \quad . \qquad [I.1.12]$$

Note that $m'_0 = 1$ and $m'_1 = \bar{x}$.

b) The r^{th} *moment about the mean* $\bar{x}$ is given by

$$m_r = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^r \quad . \qquad [I.1.13]$$

Here $m_0 = 1$, $m_1 = 0$ and $m_2 = \sigma^2$.

I.1.5.7 Coefficient of skewness

The *coefficient of skewness* is defined as the third moment about the mean divided by the third power of the standard deviation :

$$\frac{m_3}{\sigma^3} = \frac{m_3}{(m_2)^{3/2}} \quad . \qquad [I.1.14]$$

I.1.5.8. Coefficient of kurtosis

The *coefficient of kurtosis*, also known as the coefficient of pointedness, is defined as the fourth moment about the mean divided by the fourth power of the standard deviation :

$$\frac{m_4}{\sigma^4} = \frac{m_4}{m_2^2} \qquad [I.1.15]$$

I.1.5.9. Standard scores

Data are often standardised so as to be able to compare sets of data with different means and/or variances. In this case so-called *standard scores*, denoted as z_i, are used. Standard scores are defined as :

$$z_i = \frac{x_i - \bar{x}}{\sigma} \quad . \qquad [I.1.16]$$

Standard scores have an average of 0 and a variance of 1. They will be put to frequent use in the section on inferential statistics.

I.1.5.10. Grouped data

In observations $x_1,\ldots,x_N$ some numbers can be equal. Assume that the set of observations $\{x_1,\ldots,x_N\}$ is the same as the set $\{y_1,\ldots,y_p\}$ (all y_j are now different), and that y_j appears f_j times, where $j = 1,\ldots,p$, in the set $\{x_1,\ldots,x_N\}$. We then have :

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{p} f_j y_j$$

and

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{p} f_j (y_j - \bar{x})^2$$

$$= \left(\frac{1}{N} \sum_{j=1}^{p} y_j^2 f_j\right) - \bar{x}^2 \quad .$$

This follows immediately from the definitions of $\bar{x}$ and σ^2.

I.1.5.11. Other measures of dispersion

There are other measures aiming at a description of the dispersion or the concentration of data, such as the Gini index, Pratt's measure, Theil's measure and several others. These will be discussed further on in this book (see Subsection IV.7.1.3).

I.1.5.12. A graphical representation of dispersion : the box-and-whisker plot.

This graph shows selected percentiles of the data as illustrated in Fig.I.1.7 for the citation data in Table I.1.6. All values beyond the 10^{th} and the 90^{th} percentiles are graphed individually.

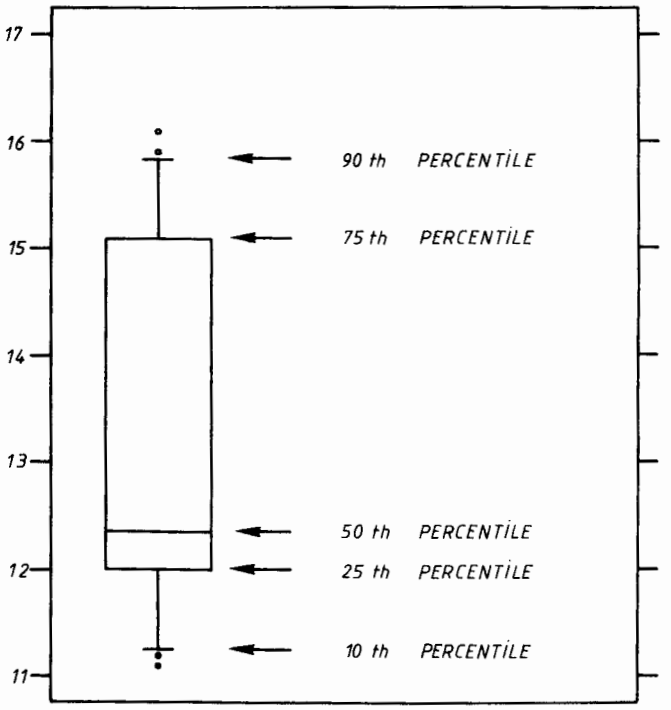

Fig.I.1.7 Box-and-whisker plot of data in Table I.1.6

I.2. ELEMENTS OF PROBABILITY THEORY

I.2.1. Probabilities

Probability theory studies situations that depend on chance. Such situations will be called 'experiments' or 'random experiments'. The set of all possible outcomes of an experiment is called the '*sample space*' or the '*universe*' of the experiment and is denoted by Ω. For instance : if the experiment is tossing a die, the sample space is $\{1,2,3,4,5,6\}$. Every subset of the universe is called an *event*. For instance, $A = \{2,4,6\}$ is the event 'an even natural number, different from zero and smaller than seven'. The probability of an event $A \subset \Omega$ is denoted as $P(A)$.

Developing the theory of probability in an axiomatic way would take up too much space and would furthermore distract us from our real objectives. Instead, we will adopt an intuitive approach and refer the reader interested in a more formal approach to books on probability theory such as Feller (1948, 1968) or Neuts (1973).

I.2.1.1. Some probabilistic equations and inequalities

(1) For every event $A \subset \Omega$: $0 \leq P(A) \leq 1$.

(2) If A^C is the complement of A with respect to Ω ($A^C = \Omega \setminus A$), then $P(A^C) = 1 - P(A)$.

(3) The impossible event, ϕ, has a probability of zero : $P(\phi) = 0$.

(4) If for every i and j, $i \neq j$: $A_i \cap A_j = \phi$, then

$$P\left(\bigcup_{i=1}^{N} A_i\right) = \sum_{i=1}^{N} P(A_i) \; . \qquad [I.2.1]$$

In particular, if $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$.

(5) If A and B are events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \; . \qquad [I.2.2]$$

(6) For any events A and B, we have

$$P(A) = P(A \cap B) + P(A \cap B^C) \; . \qquad [I.2.3]$$

I.2.1.2. Conditional probabilities

Let A and B be two events such that $P(B) > 0$. Then the *conditional probability of A given B* is denoted as $P(A|B)$ and is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad [I.2.4]$$

or

$$P(A \cap B) = P(A|B).P(B) \quad .$$

Events A and B are said to be *independent* if $P(A|B) = P(A)$ or, equivalently, $P(B|A) = P(B)$.
By [I.2.4] this is also equivalent with

$$P(A \cap B) = P(A).P(B) \qquad [I.2.5]$$

I.2.1.3. An example

We assume that the average computer time that an online system needs to find a particular paper is proportional to the number of entries in the file. When P(A) is set equal to the number of papers written by author X divided by the total number of papers in the file, we see that the time t_X needed to find any paper written by X is c/P(A), with c as a constant of proportionality.

If we know, however, that we need a paper on subject B (subject code Y), using this information to find such a paper by X will result in a probability of $P(A|B) = P(A \cap B)/P(B)$. Next, if author X writes almost exclusively on subject B, then $P(A \cap B) \approx P(A)$. Furthermore, since a subject code is usually a small part of the complete file, we get $P(B) \ll 1$. Hence $P(A|B) \gg P(A)$. So, the computer time needed to search the subfile with code Y, $t_{X \text{ in } Y}$, is much smaller than the time needed to search the whole file. Thus $t_X = c/P(A)$ and $t_{X \text{ in } Y} = c/P(A|B)$. Hence :

$$\frac{t_{X \text{ in } Y}}{t_X} = \frac{P(A)}{P(A|B)} \ll 1 \quad .$$

I.2.1.4. Bayes' rule

We finally mention, without giving the proof, the following formula, known as *Bayes' rule* : if Ω is the disjoint union of the events $A_1, A_2, \ldots, A_n$ and if, for every j, $P(A_j) \neq 0$, then, if B is an event with $P(B) > 0$:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)} \quad . \qquad [I.2.6]$$

I.2.2. Distribution functions

I.2.2.1. Discrete stochastic variables

A *discrete stochastic variable* X (also called a *random variable*) is a function from a countable universe $\Omega = \{\omega_1,\omega_2,\ldots\}$ to $\mathbb{R}$ (the real numbers) and thus :

$$X : \Omega \to \mathbb{R} : \omega \to X(\omega) \ .$$

The set $\{\omega \in \Omega | X(\omega) = x_i\}$ is an event for every $i = 1,2,\ldots$. This event is also denoted as $\{X = x_i\}$. The probability of this event is denoted as $P(X=x_i)$. The function

$$x_i \to P(X=x_i) \qquad i = 1,2,\ldots$$

is the discrete probability distribution of the stochastic variable X.

For example, let Ω be the set of all books in a library and let X be the stochastic variable which relates the book's 'age' to every book. The event $A_n = \{X=n\}$ is the set of all books that have been in the library's possession for exactly n years. The distribution of books according to age is then given by

$$\mathbb{N} \to [0,1] : n \to P(A_n) \ .$$

It is natural in this case to define $P(A_n)$ as the number of books with age n divided by the total number of books in the library.

Note that a discrete stochastic variable always satisfies the relations :

$$P(X=x_i) \geq 0 \ (i = 1,2,\ldots)$$

and

$$\sum_i P(X=x_i) = 1 \ .$$

I.2.2.2. Continuous stochastic variables

We will also use the concept of a *continuous stochastic variable*

$$X : \Omega \to \mathbb{R} \ ,$$

where Ω is a non-denumerable set. In this case $P(X=x)$, $x \in \mathbb{R}$ cannot be defined. However, such expressions as $P(x_1 \leq X \leq x_2) = P\{\omega \in \Omega | x_1 \leq X(\omega) \leq x_2\}$ are meaningful. Indeed, when variables are continuous, their individual

occurrence cannot be measured and is furthermore not important; for instance, an exact temperature, such as π degrees, cannot be measured. What can be measured is a range of temperatures $[x_1, x_2]$ (say $\pi \in [3.1, 3.2]$). $P(x_1 \leq X \leq x_2)$ then denotes the fraction of Ω that is in this situation. In mathematics, the existence of a function $f \geq 0$ is shown such that for every $x_1, x_2 \in \mathbb{R}$:

$$\int_{x_1}^{x_2} f(x)\, dx = P(x_1 \leq X \leq x_2) \quad , \qquad [I.2.7]$$

where the integral on the left denotes the area under the graph of f between the abscissae x_1 and x_2. The function f is called the '*probability density function*' of the continuous stochastic variable X. We note that density functions f satisfy

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$

I.2.2.3. Cumulative distribution functions

The *cumulative distribution* of a stochastic variable X is defined as

$$P(X \leq x) = F(x) \ , \ -\infty < x < +\infty \qquad [I.2.8]$$

If X is a discrete stochastic variable then

$$F(x) = \sum_{x_i \leq x} P(X=x_i) \quad .$$

If X is a continuous stochastic variable then

$$F(x) = \int_{-\infty}^{x} f(u)\, du \quad .$$

Conversely, we see that

$$f(x) = \frac{dF(x)}{dx} \quad , \qquad [I.2.9]$$

for continuous functions f.

I.2.3. Characteristic values of a stochastic variable

I.2.3.1. Two types of stochastic variables

a) Discrete stochastic variables

The *mean (expectation)* of a discrete stochastic variable is defined as

$$E(X) = \sum_i x_i \, P(X=x_i) \quad . \qquad [I.2.10]$$

Its *variance* is :

$$Var(X) = \sum_i (x_i - E(X))^2 \, P(X=x_i) \qquad [I.2.11]$$

$$= \sum_i x_i^2 \, P(X=x_i) - (E(X))^2 \quad .$$

If the sum in the above expressions is infinite, the mean or the variance of X is said not to exist.

b) Continuous stochastic variables

In this case the *mean* and the *variance* are defined as follows :

$$E(X) = \int_{-\infty}^{+\infty} x f(x) \, dx \quad ; \qquad [I.2.12]$$

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \, f(x) \, dx \qquad [I.2.13]$$

$$= \int_{-\infty}^{+\infty} x^2 f(x) \, dx - (E(X))^2 \quad .$$

Similar to the discrete case, E(X) or Var(X) are said not to exist when integrals do not converge .

I.2.3.2. Some theorems on the mean and the variance (no proofs are given)

1) If X is a random variable and $a,b \in \mathbb{R}$, then

$$E(aX+b) = aE(X) + b \qquad [I.2.14]$$

and

$$Var(aX+b) = a^2 \, Var(X) \quad . \qquad [I.2.15]$$

2) If X and Y are random variables, then

$$E(X+Y) = E(X) + E(Y) \ . \qquad [I.2.16]$$

3) If X and Y are independent random variables, i.e., by [I.2.5], $P(X \leq x, Y \leq y) = P(X \leq x).P(Y \leq y)$ for every x,y, then

$$Var(X+Y) = Var(X) + Var(Y) \ . \qquad [I.2.17]$$

I.2.4. Examples

I.2.4.1. The binomial distribution

Consider an experiment that can be repeated under the same conditions. Assume that this experiment has two possible outcomes : success, with probability p, and failure, with probability q = 1-p. Such an experiment is called a 'Bernoulli trial'. We are currently interested in the probability of having x successes in n independent Bernoulli trials. If X denotes the number of successes in n trials, then X is a discrete stochastic variable with values x = 0,1,2,...,n. This discrete stochastic variable has a *binomial distribution with parameters n and p* (we omit the proof) :

$$P(X=x) = \binom{n}{x} p^x q^{n-x} \qquad [I.2.18]$$

where x = 0,1,2,..., n, q = 1-p and $\binom{n}{x}$ is the binomial coefficient 'n over x', which is defined as $\frac{n!}{x!(n-x)!}$.
This is given in notation as : $X \sim B(n;p)$. If can be shown that for a binomial distribution E(X) = np and Var(X) = npq.

I.2.4.2. The Poisson distribution

Assume that patrons arrive randomly at a library's circulation desk with an average of λ arrivals per minute (or any other time unit). The probability of having n arrivals in a one-minute interval is then given by (we omit the proof) :

$$P(X=n) = \frac{e^{-\lambda}\lambda^n}{n!} \quad , \quad n = 0,1,2,... \ . \qquad [I.2.19]$$

The *Poisson law* is shown to hold for every situation in which there is a random pattern of occurrence. If X has a Poisson probability distribution, then $E(X) = \lambda$ and $Var(X) = \lambda$ (we omit the proof). This is given in notation as : $X \sim P(\lambda)$.

The property $E(X) = Var(X)$ is important for practical purposes : if, one sees in a sample that $\bar{x} \approx \sigma^2$, this is a strong indication that the property under study has a Poisson distribution. (This suspicion must, of course, be confirmed by using a statistical test (see Section I.3.5).) If an observed frequency distribution is not a Poisson distribution, the frequencies are not the result of a random process. The implication is that, in this case, it would be worthwhile to look at the situation more closely in the hope of establishing cause and effect.

Poisson's law will play an important role in queueing theory (Chapter II.3). It is also used as a model to describe multiple discoveries in science, see Simonton (1978, 1986a and 1986b) and Price (1963).

I.2.4.3. The normal distribution

A continuous stochastic variable X has a *normal distribution with parameters* μ *and* σ^2 ($-\infty < \mu < +\infty$; $0 < \sigma < +\infty$) if its density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty \qquad [I.2.20]$$

This is given in notation as : $X \sim N(\mu;\sigma^2)$.

The normal distribution is also called the 'Gaussian distribution'. If X has a normal distribution, then it can be shown that $E(X) = \mu$ and $Var(X) = \sigma^2$. For the density function $f(x)$ we note that

1) $f(x)$ is symmetric with respect to μ, and thus $f(\mu-x) = f(\mu+x)$;

2) $\lim_{x\to-\infty} f(x) = \lim_{x\to+\infty} f(x) = 0$;

3) $f(x)$ attains its maximum for $x = \mu$;

4) $f(x)$ increases for $x < \mu$ and decreases for $x > \mu$;

5) $f(x)$ has inflection points in $x = \mu \pm \sigma$.

See Fig.I.2.1.

The normal distribution is the most important probability distribution in statistics. It forms the basis for a large group of statistical tests known as parametric techniques. Several of these techniques will be discussed in Chapter I.3.

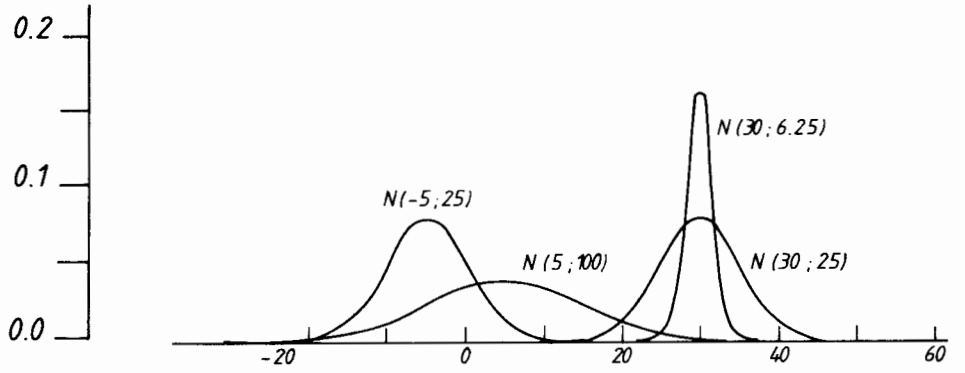

Fig.I.2.1 Various normal distributions

If $X \sim N(\mu;\sigma^2)$, then $Z = \frac{X - \mu}{\sigma} \sim N(0;1)$, as follows from [I.2.20].
The continuous stochastic variable Z is called the 'standard normal distribution' : $E(Z) = 0$ and $Var(Z) = 1$.

Values for $P(Z \leq z)$, where $z \geq 0$, are given in the appendix (Table A.1). Other values can easily be obtained from this as shown by the following calculations (based on the above properties of f) :

1) $P(Z \geq z) = 1 - P(Z \leq z)$.

2) $P(Z \leq z)$ for $z < 0$ is equal to $P(Z \geq -z) = 1 - P(Z \leq -z)$, which can be found in the tables as $-z > 0$.

3) $P(-z \leq Z \leq +z)$ (for $z \geq 0$) $= 2\,P(0 \leq Z \leq z) = 2\,(P(Z \leq z) - 0.5) = 2\,P(Z \leq z) - 1$.

4) $P(Z \leq -z \text{ or } Z \geq +z)$ (for $z \geq 0$) $= 1 - P(-z \leq Z \leq +z) = 1 - (2P(Z \leq z) - 1) = 2 - 2\,P(Z \leq z)$.

I.2.4.4. The negative binomial distribution

A discrete stochastic variable X has a *negative binomial distribution with parameters n and p* if

$$P(X=x) = \binom{n-1+x}{x} p^n q^x \quad , \qquad [I.2.21]$$

where $x = 0,1,2,\ldots$, $n > 0$, $0 < p \leq 1$, $q = 1-p$.

For this distribution $E(X) = nq/p$ and $Var(X) = nq/p^2$ (we again omit the proof). This is given in notation as : $X \sim NBD(n;p)$. The negative binomial distribution is also termed 'Pascal's distribution'. For integer values of n it gives the number of failures before the n^{th} success in a sequence of Bernoulli trials where the probability of success upon each trial is p and the probability of failure is q = 1-p. The special case where n = 1 is called a 'geometric distribution'. The NBD will play a role in library circulation models.

I.2.4.5. The negative exponential distribution

A continuous stochastic variable X has a *negative exponential distribution with parameter $b > 0$* if its density function is given by

$$f(x) = \frac{1}{b} e^{-x/b} \quad , \quad 0 \leq x < +\infty \quad . \qquad [I.2.22]$$

Then $P(X \leq x) = 1 - e^{-x/b}$. Its mean is b and its variance is b^2 (we omit the proof).

I.2.4.6. The χ^2-distribution (chi-square)

A continuous stochastic variable X has a *χ^2-distribution with n degrees of freedom* if its density function is given by

$$f(x) = 2^{-\frac{n}{2}} (\Gamma(\tfrac{n}{2}))^{-1} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \quad , \quad x \geq 0 \quad , \quad n \in \mathbb{N}_0 \qquad [I.2.23]$$

This is given in notation as : $X \sim \chi^2(n)$. The symbol $\Gamma(t)$ denotes the *gamma function*, defined as

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} \, dx \quad , \quad t > 0 \quad . \qquad [I.2.24]$$

The gamma function satisfies the recursion formula $\Gamma(t+1) = t\Gamma(t)$, and if t is a positive integer, $\Gamma(t)$ is equal to (t-1)! (t-1 factorial). We will not be

overly concerned with this gamma function as we will always use tabulated values of the χ^2-distribution.

If X has a chi-square distribution with n degrees of freedom, its mean and variance can be shown to be E(X) = n and Var(X) = 2n. We further note the following theorem (without proof). Let $X_1, X_2, \ldots, X_n$ be independent normally distributed stochastic variables with a mean of 0 and a variance of 1. Then $X = X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$. The χ^2-distribution will play an important role in hypothesis testing.

I.2.4.7. Student's t distribution

A continuous stochastic variable X has a *t-distribution with n degrees of freedom* if its density function is given by

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} , \quad -\infty < x < +\infty , \quad n \in \mathbb{N}_0 . \qquad [I.2.25]$$

This is given in notation as : $X \sim t(n)$. The t-distribution will be used in the section on hypothesis testing, where it will often play a role similar to the normal distribution, in the case of small samples. The shape of the t-distribution resembles that of the standard normal distribution. Its mean is also 0 and its variance is n/(n-2) (n > 2) (we omit the proof). As in the case of the normal distribution we also have :

$$P(-x \leq X \leq x) = 2\,P(X \leq x) - 1$$

and

$$P(X \leq -x \text{ or } X \geq x) = 2 - 2\,P(X \leq x) ,$$

where $X \sim t(n)$.

I.2.4.8. Other distributions

Other distributions will be introduced when and where they are needed. In particular, informetric phenomena are characterized by special distributions known as Lotka's, Zipf's, Bradford's, Mandelbrot's or Pareto's distributions (see Part IV). A concise statement of the foremost facts relating to the most important statistical distributions can be found, for example, in Hastings and Peacock (1975).

I.2.5. Cell occupancy problems

Many probabilistic situations can be described, regardless of whether or not they are related to informetrics, by using a model consisting of cells and objects (or balls) within the cell. For instance, a cell can represent an author and the balls in the cell papers written by this author. In this section we will mainly be concerned with three classical cell occupancies : the Bose-Einstein, the Fermi-Dirac and the Maxwell-Boltzmann distributions.

I.2.5.1. Recognisable objects

Theorem. Let $r_1,\ldots,r_n$ *be such that* $\sum_{i=1}^{N} r_i = r$. *This population of* r *objects can be distributed over the* N *cells, such that the* i*-th cell contains* r_i *objects* ($i = 1,\ldots,N$), *in the following number of ways :*

$$\frac{r!}{r_1!r_2!\ \ldots\ r_N!} \qquad [I.2.26]$$

(Bear in mind that $0! = 1$*).*

Proof. We take r_1 objects for the first cell : this can be done in $\binom{r}{r_1}$ ways. Then we take r_2 objects for the second cell : this can be done in $\binom{r-r_1}{r_2}$ ways. We continue in this manner until we reach the last cell for which there is no longer any choice : we have to take the remaining r_N objects. This yields a total of

$$\binom{r}{r_1}.\binom{r-r_1}{r_2} \cdot \ldots \cdot \binom{r - \sum_{j=1}^{N-1} r_j}{r_N}$$

choices. According to the definition of the binomial coefficients, this is equal to :

$$\frac{r!}{r_1!(r-r_1)!} \cdot \frac{(r-r_1)!}{r_2!(r - \sum_{j=1}^{2} r_j)!} \cdot \ldots \cdot \frac{(r - \sum_{j=1}^{N-1} r_j)!}{r_N!\ 0!} = \frac{r!}{r_1!r_2!\ \ldots\ r_N!} \qquad \square$$

If the r_i are not fixed, there are N^r possible cell occupancies (i.e. the number of mappings from a set of r elements to a set of N elements : for every element in the first set, there are N possible assignments). The *Maxwell-Boltzmann distribution* assumes that all these N^r cases are of equal probability,

so that each has a probability of $1/N^r$.

The probability of obtaining a specific cell occupancy (r_i's fixed) then becomes :

$$\frac{r!}{r_1! \ldots r_N!} \cdot N^{-r} \quad . \qquad [I.2.27]$$

Based on the above reasoning, the Maxwell-Boltzmann distribution would seem to be the most logical one. However, less intuitive distributions are more commonly encountered (especially in physics). These will be studied in Subsection I.2.5.2.

I.2.5.2. Unrecognisable objects

In the case of unrecognisable objects, only the number of objects in every cell is important. Switching two objects from different cells leaves the distribution unchanged. If a total of r objects is to be distributed over N cells, every solution (i.e. every N-tuple) of the equation

$$\sum_{i=1}^{N} r_i = r \ , \quad r_i \geq 0 \ ,$$

yields a possible configuration. Two cell distributions can be distinguished if the corresponding N-tuples $(r_1, r_2, \ldots, r_N)$ are different.

Theorem. 1. The number of distinguishable cell distributions (r objects over N cells) equals

$$A_{r,N} = \binom{N+r-1}{r} = \binom{N+r-1}{N-1} \quad . \qquad [I.2.28]$$

2. The number of distinguishable cell distributions in which no cell remains empty is :

$$B_{r,N} = \binom{r-1}{N-1} = \binom{r-1}{r-N} \quad . \qquad [I.2.29]$$

Proof. 1. Let us visualise a configuration of N cells with r objects as a row of bars and stars such as in the following configuration :

|**|*|||***|*||

There are consequently N+1 bars and r stars. Every configuration is obtained by placing r stars in (N+1+r) - 2 = N+r-1 spaces (begin and end with a bar).

This can be done in

$$\binom{N+r-1}{r}$$

possible ways. Once the stars have been assigned, the remaining spaces are automatically occupied by the N-1 bars. This proves part 1 of the theorem.

2. Demanding that no cell should remain empty is equivalent to requiring that no two bars should be adjacent. Let us therefore consider a row of r stars and observe that there are r-1 spaces between the stars. Let us next pick N-1 of these r-1 spaces (this is only possible if $N \leq r$) for the bars. This can be done in $\binom{r-1}{N-1}$ different ways. This ends the proof of the theorem. □

The numbers $1/A_{r,N}$, expressing the equal probability of all distinguishable cell distributions, yield the so-called *Bose-Einstein distribution*. It is often applied in the theory of photons, nuclei and atoms containing an even number of elementary particles, see e.g. Feller (1968, p.41).

The *Fermi-Dirac distribution* assumes that :

a) it is impossible for two or more particles to be in the same cell (hence $r \leq N$ and for every $i = 1,\ldots,N : r_i = 0$ or 1);
b) all distinguishable distributions satisfying a) have an equal probability of occurring.

Hence, with the Fermi-Dirac distribution, there is a total of $\binom{N}{r}$ possible arrangements, each having probability

$$\binom{N}{r}^{-1} \qquad [I.2.30]$$

This model applies not only to electrons, neutrons and protons, but also to misprints in a book. If a book contains N symbols of which r are misprints, this situation can be configured as N cells and r balls, with at most one ball in each cell. The distribution of misprints consequently follows a Fermi-Dirac distribution.

Another theoretical application of the use of the Fermi-Dirac occupancy model is in the use of files of library patrons. When entering the library for the first time, users fill out a form stating their name, date of birth, address and so on. This form is all that remains in the library once the user has left. Especially when forms are incomplete, one might ask if every library user is uniquely determined by these data. If yes, the data follow a Fermi-Dirac distribution in the total population of the region served by the library. For more information on this topic, the reader is referred to Leiser (1972).

We will conclude this section on occupancy problems by applying the second part of the above theorem. Assume that we are observing the entrance to a library. If the person entering is a male, we write down the letter M and if it is a female we write down the letter F. After a certain amount of time, we obtain a chain such as :

MMFFFMFMMMMFFMFFF .

Suppose further that we terminate our observations after a certain time. The following questions may then be posed :

(i) What is the probability of obtaining a distribution of M's and F's as observed?

(ii) Do persons enter the library in groups of the same sex (a question on human social behaviour)?

Without performing any statistical test we can already solve part of these questions. Suppose we have observed m men and f women; suppose moreover that we have n runs of M's (i.e. groups of consecutive M's). Hence there are n-1, n (as in the above chain) or n+1 runs of F's. For illustrative purposes, let us say that there are n+1 runs of F's. Consequently, n runs of M's means, in fact, that n cells are to be occupied by M's, but no cell may remain empty. According to the above theorem, this can be done in $\binom{m-1}{n-1}$ different ways. Analogously, the n+1 runs of F's yield $\binom{f-1}{n}$ possibilities. Altogether, there are $\binom{m-1}{n-1}\binom{f-1}{n}$ possibilities for n runs of M's and n+1 runs of F's. The total number of different situations, with every number of runs allowed, is clearly

$$\binom{m+f}{m} = \frac{(m+f)!}{m!f!} = \binom{m+f}{f} \quad .$$

Hence, question (i) boils down to computing

$$\frac{\binom{m-1}{n-1}\binom{f-1}{n}}{\binom{m+f}{f}} \quad . \qquad \text{[I.2.31]}$$

For n = 1,2,..., min(m,f), the above equation yields a discrete distribution, as shown by Fig. I.2.2.

As can be seen from this figure, the points belong to a more or less normal distribution and it can be shown that, in the limit, this is indeed the case. On the left we encounter the situation n = 1, i.e. MMM...MFF...F (perfect clustering) and on the right the situation n = min(m,f) MFMFMF...MF (anti-clustering), with the random situation in the middle.

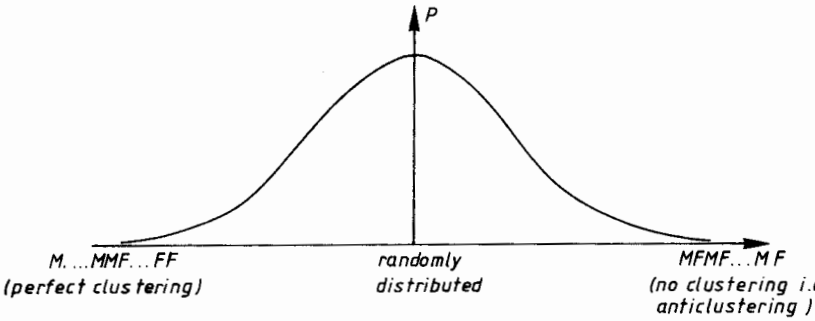

Fig.I.2.2 Distribution of M's and F's

Based on these considerations, it is possible to test hypotheses to find out whether or not arrivals occur in groups according to sex. The Wald-Wolfowitz runs test (see Subsection I.3.7.2) can be used for this purpose. Note that if only the left tail is rejected we can conclude that the arrivals are distributed equally; if both the left and the right tails are rejected, we can conclude that the arrivals are distributed randomly.

The Bose-Einstein distribution has been used by Hill (1974) to derive an important informetric law (to be discussed in Part IV), namely Zipf's law. Hill's procedure was also applied later in Orlov et al. (1985), where thermodynamic principles were used to describe document distributions.