5

The pragmatics of information retrieval experimentation

Jean M. Tague

The novice information scientist, though he or she may have thoroughly studied the design and results of previous information retrieval tests and clearly described the purpose of his/her own test, may still, when faced with its implementation, have great difficulty in proceeding. Early information retrieval experiments were of necessity ad hoc, and it is only in recent years that a body of practice, based on the experiences of Cleverdon and later investigators, has made possible a few recommendations on the pragmatics of conducting information retrieval experiments.

The following remarks, though based to some extent on a study of the major tests, including those described in later chapters of this book, are heavily dependent on the author's own trials, tribulations, and mistakes. If there is one lesson to be learned from experience, it is that the theoretically optimum design can never be achieved, and the art of information retrieval experimentation is to make the compromises that will least detract from the usefulness of the results.

In determining experimental procedures, three aspects must be kept in mind:

(1) The validity of the procedure; does it determine what the experimenter wishes to determine? If a study is being made of the relation of document scope to user satisfaction, does the use of number of citations as a measure of scope and number of references marked 'relevant' as a measure of satisfaction really fulfill this purpose?
(2) The reliability of the procedure; can it be replicated by another experimenter? If one is addressing the problem of inter-indexer consistency, will a test of the consistency of two indexers indexing 10 documents from a single journal provide results which can be replicated elsewhere? A procedure may be reliable without being valid, i.e. it may give consistent results but be measuring something else.
(3) The efficiency of the test procedures; how long will it take, how many resources—people, computing, supplies, equipment—will it require, how much will it cost? Is it sensible, for example, to assess the absolute recall of searches when this means each user will have to peruse the entire database? What limitations will this place on the size of the database?

The approach to information retrieval testing in this chapter will be to step through an information retrieval test procedure, indicating, at each step, the choices that will face the experimenter. Suggestions will be made for resolving these in ways that take into account the validity, reliability, and efficiency of the experiment. It is assumed that the experimenter has decided what is to be tested, bearing in mind the problems discussed in the three preceding chapters, and can clearly distinguish this from the assumptions she/he is making.

5.1 Decision 1: To test or not to test?

It should be unnecessary to point out to information scientists the necessity of a thorough literature search before embarking on any experimentation at all. Unfortunately, even in this field one finds attempts to reinvent the wheel. *Library and Information Science Abstracts*, the *Annual Review of Information Science and Technology*, *Information Science Abstracts*, *Library Literature*, *Computing Reviews*, *Computer and Control Abstracts*, *Dissertation Abstracts* are required reading prior to planning. Although the actual experiment may not have been attempted previously, some partial or suggestive results may be available. Previous papers frequently bring to the attention of the investigator useful methodology or even sets of queries and evaluations. Many writers have pointed out the need for cumulative studies in information retrieval. Only a thorough grounding in previous research will make this possible.

5.2 Decision 2: What kind of test?

This decision relates to the broad category of test. Will it be a laboratory or operational test? Will it be a complete or partial test? Cleverdon[1] made these distinctions clearly in Cranfield 2. Operational tests normally involve an evaluation of an existing system, laboratory tests attempt to advance knowledge about individual variables of information retrieval. Complete tests involve three aspects: a collection of documents, a set of search requests, and relevance judgements relating the search request to the documents. Partial tests, as the name implies, usually are concerned with aspects of the document set other than retrieval.

The characteristics of the various types of test are discussed in more detail in Chapter 2. Although the purpose of the test will influence the choice of laboratory versus operational system, other factors must be considered. If a laboratory test is to be mounted, does the experimenter have the time, the people, and the funding to carry it out? Laboratory tests tend to be more expensive than operational tests of the same size. On the other hand, tests of operational systems should not be attempted unless there is an assurance of co-operation from the operational personnel, not just top-level management. In this regard, face-to-face conversation or at least telephone contact is more effective than written correspondence.

5.3 Decision 3: How to operationalize the variables?

A variable is simply some attribute or feature—qualitative or quantitative—of a retrieval system. In experimental work, variables are normally classified as independent or input or system variables, on the one hand, and dependent or output or performance variables, on the other hand. The independent variables are the ones the experimenter manipulates or controls in order to determine the effect on the dependent variables—for example, effect of indexing depth on indexing speed; effect of term linkages on recall and precision. It must be remembered that there are no *a priori* independent and dependent variables. An independent variable in one experiment may serve as a dependent variable in another. For example, one study might be of the effect of different indexing language features on speed of indexing, another on the effect of speed of indexing on the number of indexing errors.

The three previous chapters have discussed many of the independent and dependent variables of previous information retrieval experiments. This chapter will be concerned solely with procedures for actually observing or measuring them, i.e. with operationalizing them. Because of the great variety of variables encountered in information retrieval—variables characterizing the document collection and database, the indexing languages, the queries and search processes, the people associated with a retrieval operation, and the evaluation of the output—discussion will be retricted to those which have been previously operationalized. This is not to suggest that these operationalizations are necessarily the best or only ones. However, the approaches may be useful to the new information scientist.

What follows is in the form of a listing of the major categories of variables, with some suggestions for operationalization.

Document collection and database

Variables in this category relate to the size, source, form, medium, and broad subject coverage of the collection and to the type of document representation or surrogate used in the database. The term database here will refer to the collection of document surrogates and any associated indexes or access files. The term is used very generally and includes printed indexes, card catalogues, microform indexes and catalogues, and tape or disk files, whether accessed in batch or interactive mode. Since most of these variables are qualitative, the main problem here is to define appropriate categories. What are the possible document forms—monograph, journal article, technical report, patent, etc., and to what extent can they be applied across the different media—print, micromedia, film, recordings, machine-readable text? What are the possible elements in a database record—author, title, source, index terms, abstract, full text, etc.?

An important variable here is the heterogeneity of the collection. One approach, presented by Brookes[2], is a measure of categorical dispersion. The documents are assigned to n subject categories. The categories are ranked and the frequency in each category is determined. If $f(r)$ represents the frequency of the rth category, the mean rank will be

$$m = \frac{\sum_r rf(r)}{\sum_r f(r)}$$

and a measure of dispersion which varies between 0 and 1 is

$$D = \frac{2(m-1)}{n-1}$$

Brookes has modified this measure to take into account the relative sizes in the total population.

Indexing languages and indexing procedures

Many investigators have contributed to the operationalization of indexing variables. Below are listed some of the more generally accepted definitions.

(1) Exhaustivity of indexing, i.e. the number of topics covered by the indexing. Operational definition: number of index terms/document. Keen and Wheatley[3] suggest redundant indexing, for example synonyms and morphological variants, be eliminated.
(2) Specificity of indexing, i.e. the preciseness of the subject description. Operational definition: number of postings per term. This value, however, may depend as much on fashions in the literature as on the specificity of a term.
(3) Degree of linkage in a vocabulary. Operational definition: number of references in the dictionary or thesaurus. Keen suggests that only *see also* references, not *see* references, be included.
(4) Degree of vocabulary control. Operational definition: number of terms in the entry vocabulary/number of terms in the indexing vocabulary. If indexing is uncontrolled this value becomes 1.
(5) Term discrimination value. Operational definition (Salton, Wong and Yu[4]): Document surrogates are vectors (binary or weighted) of index terms. The similarity between two documents

$$d_1 = (d_{11}, d_{12}, \ldots, d_{1n})$$
$$d_2 = (d_{21}, d_{22}, \ldots, d_{2n})$$

is measured by the cosine coefficient:

$$S(d_1, d_2) = \frac{\sum_{j=1}^{n} d_{1j}\, d_{2j}}{(\sum_{j=1}^{n} d_{1j}^2 \sum_{j=1}^{n} d_{2j}^2)^{1/2}}$$

The centroid of a set of documents is

$$C = (c_1, c_2, \ldots, c_n)$$

where

$$c_j = \frac{\sum_{i=1}^{m} d_{ij}}{m}$$

The summation is over all m documents in the set. The term discrimination value of the jth term is then

$$D_j = \frac{Q_j - Q}{Q}$$

where Q, the compactness of the collection, is defined as the average similarity of the documents with the centroid

$$Q = \frac{1}{m} \sum_{i=1}^{m} S(c, d_i)$$

and Q_j is the compactness of the collection when term j is deleted:

$$Q_j = \frac{1}{m} \sum_{i=1}^{m} S^*(c, d_i)$$

where

$$S^*(c, d_i) = \frac{\sum_{k \neq j} c_k d_{ik}}{(\sum_{k \neq j} c_k^2 \sum_{k \neq j} d_{ik}^2)^{1/2}}$$

Thus, a term is discriminating to the extent that it decreases the average similarity of the document set.

(6) Degree of pre-coordination of index terms. Operational definition: number of index terms per index phrase. Averaging may take place over either all entries (types) in the dictionary of the language or over all tokens in the database.

(7) Degree of syntactic control (i.e. roles, links, relational operators). Operational definition: number of operators per documents. Since this measure confounds indexing exhaustivity and syntactic control, a better measure might be the ratio: number of operator assignments/number of index term assignments.

(8) Accuracy of indexing. Operational definition: number of indexing errors, as determined by a judge or by reference to a standard set of term assignments. Two types of errors are distinguished: of omission (a term omitted) and of commission (an incorrect term added). Validity presents a problem here: why is the judge or standard more 'correct'? It is difficult to validly assess indexing correctness without retrieval. Retrieval, however, is no real solution. Why should a particular set of queries be used to test indexing? No indexer or judge can foretell all future uses of a document. The best one can do is assume that the best judges will be those who have worked with the users of the collection.

(9) Inter-indexer consistency. Operational definition: various ratios lying between 0 and 1 have been suggested, for example

$$\frac{N(\mathrm{A} \cap \mathrm{B})}{(N(\mathrm{A})N(\mathrm{B}))^{1/2}} \quad \text{or} \quad \frac{N(\mathrm{A} \cap \mathrm{B})}{N(\mathrm{A} \cup \mathrm{B})}$$

where $N(\mathrm{A})$ and $N(\mathrm{B})$ are the numbers of index terms assigned by indexers A and B, respectively, $N(\mathrm{A} \cap \mathrm{B})$ is the number of terms assigned by both A and B, and $\mathrm{N}(\mathrm{A} \cup \mathrm{B})$ is the number of terms assigned by either A or B.

Queries, search statements, and the search process

'Query' will be used here to mean the verbalized statement of a user's need. A 'search statement' is a single string, expressed in the language of the system, which triggers a search of the database, i.e. causes a search algorithm to scan the database and output a response. A 'search process' is a sequence of search statements, all relating to the same query.

Search statements can be categorized by type. Some types are:

single search elements
boolean combinations of search elements
binary vectors of search elements
weighted vectors of search elements
any of the above with syntactic requirements:
    roles
    facets
    word adjacency
    word dependency
    co-occurrence frequency

The notions of specificity and exhaustivity can also be applied to search statements, but here the development of operational measures must take into account that search statements are often boolean expressions, i.e. combinations joined by the connectives and, or, not (called conjunction, disjunction, and negation). Measures of specificity and exhaustivity generally require that an expression be put into a standard form, for example into disjunctive normal form, as a disjunction of conjuncts. For example, the following is in disjunctive normal form

$$(T_1 \wedge T_2 \wedge T_3) \vee (T_1 \wedge T_2 \wedge T_4) \vee (T_4 \wedge \neg T_5)$$

but the following equivalent statement is not

$$T_1 \wedge T_2 \wedge (T_3 \vee T_4) \vee \neg\{\neg T_4 \vee T_5\}$$

A measure of the breadth or exhaustivity of the search is the number of or's ($\vee$); of the specificity the average number of and's ($\wedge$) per conjunct. In the example above, the breadth is 3 and the depth is 2.67.

A problem arises in applying these measures when the system permits truncation of search terms, as do most of the commercial online systems. In this case, one must refer to the dictionary, if it exists, to determine how many discrete terms correspond to a given truncation. With no dictionary, it may be possible to determine this number by a search of the database.

Other measures relate a document representation and a search statement. The simplest of these is co-ordination level or degree of match, counting the number of terms the document and query have in common. The measure may be normalized in a manner similar to the normalization of inter-indexer consistency, i.e.

$$\frac{N(q \wedge d)}{\sqrt{N(q)N(d)}} \quad \text{or} \quad \frac{N(q \wedge d)}{N(q \vee d)}$$

where the search statement q has $N(q)$ terms, the document d $N(d)$ terms, and $N(q \wedge d)$ terms match. Salton's cosine coefficient, which has the same form as the document–document cosine coefficient previously defined, assumes a vector representation of both document and query and is a more general measure of document–query similarity.

The action a searcher takes in response to a query is, in most general terms, a sequence of search statements. It is only in searching computer databases in batch mode that the single search statement is the norm. In searching

printed indexes, card catalogues, and other manual files and in online interactive computer searching, the response to one search statement triggers the next, until the searcher is satisfied.

The length of this process can be measured by the total number of search statements or the search time. These measures, however, confound the complexity of the search and the experience or style of the searcher. Variables describing overall complexity as a function of inter-statement relationships need to be developed.

Another search process variable is the form of the output: citation, citation and abstract, citation and index terms, citation and full text, etc. Particularly in online systems, there may be a choice of output format. In fact, in online searching the term output is ambiguous. Typically, one keys in a boolean search statement and gets as response a statement of the number of hits. If this is large, no further output at all may be requested, or a few titles only may be scanned. The strategy may then branch in one of several directions:

list all citations online
modify the search statement
list all citations offline
terminate the search

In an experimental situation, a choice must be made as to what constitutes output. No consensus appears to have yet been reached in the field. This is unfortunate, as measures of effectiveness obviously depend on how output is defined.

Marcus, Kugel and Benenfeld[5] have introduced the idea of the indicativity of an output field. It is defined operationally as the proportion of documents which the field indicates to be relevant which are actually assessed as relevant from full text. It appears that indicativity is related logarithmically to the length (i.e. number of words) in the field.

People—indexers, searchers, users

Because many of the decisions in information retrieval are based on human judgement, the professional training and experience of the system personnel and of the users is often studied. These variables can usually be operationalized in terms of years of experience and/or training, in general or with particular systems. Job titles are an unreliable guide, as they vary from installation to installation. Number of previous searches can also be used as a measure of experience with a particular system.

Many of the things one wants to know about people can be characterized as attitudinal variables: ease with other people, acceptance of automated processing. Experimenters should be aware of the large psychological literature on attitude measurement, and consider the use of one of the standard scales—Guttman, Lickert, Thurstone, the 'unfolding' method.

A good text is Lemon[6]. He defines attitude measurement as gathering observations about people's behaviour and allocating numbers to these observations according to certain rules. Attitude scales depend on the investigator's theoretical assumptions about the nature of the attitude he or she is trying to measure, its relationship with behaviour, and the rules used

to assign numbers. He points out that there is nothing sacrosanct about the well-known scales, and they can be suitably modified by the investigator.

The interaction between client or user and search analyst or reference librarian known as query negotiation is being increasingly studied. Zipperer[7] has analysed this interaction in terms of nine activity categories;

Question negotiation (presentation of query)
Profile development (vocabulary selection)
Tutorial activities (explanation)
Search type selection (current awareness or retrospective)
Strategy formulation (search statement specification)
System description
Database selection
Administrative procedures
Diversionary activities (interruptions)

Although these activities relate more to batch than online retrieval and thus might be modified for an interactive environment, it is important that this kind of analysis be standardized, so that results from different studies may be compared.

Evaluation

Historically, the 'evidence' of information retrieval experiments has been in the form of retrieval effectiveness measures, and more specifically recall and precision. Cleverdon[1] pointed out the reason for this continuing popularity of these two measures:

> 'The unarguable fact, however, is that they are fundamental requirements of the users, and it is quite unrealistic to try to measure how effectively a system or subsystem is operating without bringing in recall and precision.'

How one calculates recall and precision depends on the ordering of the output. There are four possibilities:

Unordered output, i.e. output is the retrieved set.
Ranked output, with possible ties in ranking.
Totally ranked output, i.e. each document has a unique rank.
Weighted output, i.e. each document has a weight.

If the retrieval set is unordered, then a four-way partition of the full database is made to determine recall and precision:

a is the number of relevant and retrieved references.
b is the number of non-relevant and retrieved references.
c is the number of relevant and non-retrieved references.
d is the number of non-relevant and non-retrieved references.
$n = a+b+c+d$ is the total number of references in the database.

Four measures have been defined in terms of this partition:

recall $= a/(a+c)$
precision $= a/(a+b)$
fallout $= b/(b+d)$
generality $= (a+c)/(a+b+c+d)$

Any three of these will determine the fourth. For example, if recall = 1/2, generality = 2/27, fallout = 1/100, then

$a = c$, $d = 99b$, $27(a+c) = 2(a+b+c+d)$, $a = 4b$ and
precision = 4/5 = 0.8.

However, two alone do not determine the others. For example, if generality remains constant at 2/27, but recall increases to 3/5, then precision may increase, remain constant, or decrease depending on fallout. If fallout remains constant at 1/100, then

precision = 24/29 = 0.828.

If fallout increases to 1/50, then

precision = 12/17 = 0.706.

If fallout increases only to 3/250, then

precision = 4/5 = 0.8

i.e. remains constant. Thus, the statement that as recall increases, precision decreases, may be an empirical characteristic of a particular retrieval system, but does not follow formally from the properties of recall and precision.

If output is ranked, totally or with ties, then recall and precision can be calculated at each rank, using the rank as a retrieval threshold. If output is weighted, recall and precision can be calculated at standard weight thresholds. In both cases, values may be averaged to obtain a single value. However, this approach is not very realistic, as all threshold levels are not equally likely to be appropriate for a query. Some form of weighted averaging may be more appropriate.

Two practical problems arise in determining recall and precision:

How is relevance of the references to be assessed?
How are all relevant items in the file to be found?

A thorough review of the concept of relevance will be found in Saracevic[8]. Pragmatically, the problem lies in deciding on the scale of relevance and then instructing the evaluators so that they will carry out the relevance assessments in a consistent manner. The following scales have been used:

Binary relevance: a reference is either relevant or non-relevant.
Three-value relevance: a reference may be relevant or highly relevant, probably or partially relevant, or non-relevant.
Ranked relevance: references are ranked with respect to relevance. Ties may or may not be permitted.
Relevance weights: each reference is assigned a weight by the user, indicating the strength of its relevance to the query.

In choosing among these scales, consideration must be given to reliability, i.e. is there consistency in the relevance ratings by the same individuals at different times and different individuals for the same query? Studies, for example Lesk and Salton[9] and Rees and Schultz[10], have shown relevance rankings to be relatively stable. Borderline problems frequently arise in making a binary distinction, and these are not really solved by the three-value scale. This simply replaces one borderline by two. Relevance weights

have been little used, perhaps in recognition of their inherent unreliability. Most psychologists use no more than seven points in a scale, perhaps because it has been found that humans can rarely make distinctions beyond this range. (See Miller[11], for example.)

It is important that all individuals making relevance assessments receive the same instructions. It has frequently been pointed out that relevance embodies two distinct notions:

> Is the document an answer, i.e. is it about the subject of the query?
> Will it be useful to the user? If, for example, the user has already read the document, it will not be useful?

Users should be clear whether they are assessing subject relevance or pertinence. Sometimes pertinence is operationalized by asking the question:

> Would you look at the document represented by this reference?

Or by checking whether, in fact, the user did order or read the document.

In addition, users should be instructed what form of output should be used in making relevance judgements. Previous experiments have shown (see Saracevic[8]) that relevance judgements are influenced by form: title, full citation, abstract, full text.

Determination of the full set of relevant documents in the file, which is necessary for determining recall, is a problem which has dogged information retrieval experimentation since Cranfield 1. Some solutions which have been used are as follows:

(1) One or more predetermined relevant documents are included in the file. The problem here is that unless the full file is perused, one cannot be sure other documents may not be relevant. Two ways of predetermining the relevant set are (a) asking the author of a paper to state a query based on the paper and assess the relevance of all papers cited, and (b) use the title as a query and the cited papers as the relevant set. This second approach, in particular, has the disadvantage that relevance is operationalized in a very arbitrary, non-judgemental fashion and hence is of questionable validity.

(2) Use a small document set and have the relevance of all documents for all queries assessed by users or system personnel. Here, of course, the problem is that small files are not very reliable, i.e. they are subject to wide variation from file to file.

(3) Take a random sample from the file and assess all documents in the sample as to relevance. The problem here is similar to that with small files. In most operational systems, the generality will be very low, so that the sample size needed to assess it accurately will be very large. For example, if there are actually 50 relevant documents in a file of 50 000 (a not unreasonable generality of 1/1000), then the sample size needed to estimate the total number of relevant documents at a 95 per cent confidence level and error less than 0.0001 will be the value of n which satisfies the equation

$$0.0001 = 1.96\sqrt{\frac{0.001(0.999)}{n}}\sqrt{\frac{50\,000-n}{50\,000}}$$

i.e. $n = 44\,237$.

A detailed analysis of the percentage of a pool of documents which must be assessed in order to test statistically for a difference between two methods at specified significance and power levels is given in Gilbert and Sparck Jones[12]. A second method gives the actual numbers of documents.

(4) In comparative tests, instead of calculating absolute recall calculate relative recall. This is defined as follows:

Let A_i, $i = 1,m$ be the set of relevant documents retrieved by the ith treatment or level of the variable. Then the relative recall of the ith system is defined by

$$\frac{|A_i|}{|\bigcup_{i=1}^{m} A_i|}$$

where $|X|$ represents the number of elements in the set X. So the recall of the ith treatment or level becomes the proportion of relevant documents retrieved by any system which are retrieved by the ith treatment or level.

Relative recall seems appropriate in comparative testing, though it obviously cannot be used to compare results from one experiment or database to another. The values are heavily dependent on the particular treatments under consideration. It is virtually the only possible approach to recall in testing large operational systems.

Evaluation also considers variables relating to efficiency—time, cost, cost/benefit, cost/effectiveness. Although times such as searching time or document delivery time or total response time (i.e. time between the first and final contact of the user with a system) present no conceptual difficulties, in practice, with operational systems, the values are difficult to collect. Computer systems usually provide information about connect time (i.e. elapsed time) and CPU time (time the computer was actually processing data). Problems may arise with computer down time. Frequently, when the system crashes, no record will remain of time already spent on the system (or money either, which may be an economic advantage but an experimental problem). If system crashes are frequent with online systems, searchers are advised to keep their own time records as well.

Paralleling computer crashes is the problem of interruptions in manual searching. If these are more than remote possibilities, then each searcher rather than a single time keeper should keep time records.

Costing a retrieval system, overall and for individual searches, is not a trivial undertaking. Such costs must include:

Personnel time—professional and clerical
Communication time
Equipment costs, suitably amortized
Supplies
Document reproduction
System overhead—rent, utilities, taxes, etc.

among other items. Sometimes the cost to the user of the time he or she spends interacting with the system is also included.

Obtaining cost data requires meticulous record-keeping by the staff. This is not always an accepted practice in operational systems, and the

experimenter may have to develop procedures for obtaining cost data and persuade system personnel to carry them out. Always investigate what cost information is routinely collected before beginning an experiment. Don't expect that procedures will necessarily be changed to suit your needs. Persuasion, charm, and bribery may be required.

Cost effectiveness and cost benefit are really two distinct concepts. The former relates the cost of a retrieval system to its effectiveness in serving its users. Cooper[13] has suggested the following measures of cost effectiveness:

$C1$ = cost/retrieved reference
$C2$ = cost/relevant reference
$C3$ = cost/precision
$C4 = C2 - C1$.

Cost/benefit relates the cost of a system to the overall benefit it provides within a society or community or institution. Defining social benefit operationally, rather than simply assessing its importance, is an idea whose time has not yet come in information retrieval.

It must be emphasized that operationalizations have been cited in this section purely as examples, not in any sense as the only valid definitions. Other ways may have equal or greater validity depending on the purpose and environment of the experiment.

5.4 Decision 4: What database to use?

There are three alternatives here, each with its own advantages and disadvantages: (1) Build an experimental database; (2) Use an existing experimental database; (3) Use an operational database.

Building your own database is expensive, so that, unless the investigation is lavishly funded, it will necessarily be small. There is little evidence that, in information retrieval, one can extrapolate findings from small databases to large ones.

The size of an experimental database is a much-debated problem. Test collections surveyed by Sparck Jones and Van Rijsbergen[14] ranged in size from 300 to 50 000. However, the larger databases were normally derived from operational databases and/or used derived (e.g. from title) rather than assigned indexing. The authors suggest that research needs appear to be for operationally-derived collections of 30 000 documents, with subcollections of 2000 having special properties. Very little is known about the variability of recall and precision under varying collection size. Tague and Farradane[15] showed that the sampling error in estimating system recall and precision from samples is inversely proportional to the square root of the collection size (see Section 5.9).

Experimental databases, either self-constructed or obtained from previous experiments, are almost essential in comparative indexing studies. Only then is it possible to exercise the necessary control. Many different kinds of control are needed, among them control of the collection coverage, the form of surrogate, the characteristics of the indexing. These will be discussed individually.

Collection coverage

The subjects of the documents described by a database, their age, language, scope, medium and source can all, potentially, affect the measures of retrieval performance. Hence, one must either use a collection which is homogeneous with respect to these attributes and then claim results only for this limited sphere or attempt to randomize the collection with respect to some or all of them. Some form of randomized selection, even within a narrow boundary, is essential. This eliminates a possibly unconscious bias of the experimenter in selecting the documents. For example, if documents were to be from the computer science field and published during the past three years in English, a random selection from an existing bibliographic database such as *Computing Reviews* or *Computer and Control Abstracts* could be used. Tables of random numbers are useful in making the selection, either to select document numbers if items are numbered or to select pages and line number within the page if they are not.

Form of surrogate

The form of the document surrogates, whether citations only or citations with index terms, abstracts, full text, etc., should be appropriate to the hypothesis under test. Also, form of output presented to a user affects relevance judgements. It is essential, therefore, that all entries in the database be in the same form. Also, if real users, with real information needs, are involved in the experiment, there should be access to the full text of the documents themselves, if only for 'public relations'.

One might not wish, however, to make decisions about record form solely on the basis of present needs. Because of the expense of setting up an experimental database, consideration should be given to future use of it, both by the investigator and by others. If additional fields can be input at very little added cost, fields which have a high probability of being useful in later experiments, it often saves time to include them, particularly if short. Also, it is useful to include one or two blank fields in a computer record, which can be assigned later.

Characteristics of the indexing

If documents are indexed using a number of different languages, how will the investigator ensure that parallel index records in different languages cover the same topics? Keen[3] has described the use of an intermediate language, into which all topics to be indexed are initially described, for this purpose. Other aspects of the indexing process which should be controlled are the professional level and experience of the indexers, and the source of the indexing, whether from full text, abstract, or title. It is an obviously biased procedure to use the same personnel for both indexing and searching. In addition, one would prefer to see the chief investigator in a study remain relatively independent of both these operations. However, for research with a small budget, as, for example, much doctoral research, this requirement may simply be impossible to satisfy.

The structure of the database should be appropriate to the type of query

that will be processed. The norm in computer-based information retrieval is a file of document surrogates in random order or ordered on some semi-random attribute such as accession number, with one or more associated inverted files to access by index term, author, title term, abstract term, or other aspect of interest to the investigator. The advantage of the random sequence is that documents can be added to the file without reorganizing it.

Before setting up inverted file indexes to a set of document records, a number of choices must be made:

(1) What attributes or fields will be indexed?
(2) Will the indexing be based on the complete string within a field or on individual words within the field?
(3) Will all individual words be indexed or will there be a stop list?
(4) Will words be stemmed?

These choices will be dependent on the purpose of the experiment, but the wise investigator will think out all implications before setting up the database. For example, is stemming economically justifiable if a truncation operation can be used in searching?

Some experimental databases, notably the Smart system, use a clustered organization. This structure often increases search efficiency and reduces search time. Of course, there is processing time involved in the original clustering, but if many searches are processed, there may be an over-all benefit. More efficient clustering algorithms are constantly being developed, so that if one intends to follow this route, a survey of recent computer science literature would be in order. A good survey of clustering algorithms is given in Hartigan[16]. A simple single link algorithm is described in Salton[17].

The medium of the database—whether it is computer-based or microform or printed—and, if computer-based, whether it will be accessed in batch or online mode, is a decision that will usually be made in the early stages of a project, because of its implications for the resources which will be required. Sometimes, the choice will be predetermined by the nature of the experiment or the availability of facilities. Where there is a choice, the investigator should consider the following points:

(1) At the data entry stage, computer-based files are more efficient, as each document record needs to be keyed once only.
(2) Corrections, reformating for printed output, sequencing for storage, production of multiple printed records can all be carried out automatically with machine-readable input. Even essentially manual files such as card catalogues are now produced by computer. Word processing equipment is useful in generating small printed files.
(3) In-house computer files require a set of programs for the initial set-up of the database, for maintenance, and for retrieval. (See the next section for further comments on the design of these.)
(4) Online systems offer much greater flexibility in searching and in analysing searches.
(5) The cost differential between online and batch is rapidly changing in favour of online.

Online retrieval is rapidly becoming the norm in libraries, businesses, and scientific institutions. It seems inevitable that the information retrieval field

will continue to move in this direction because of cost reductions and advances in communication technology. For small databases, an online system is strongly recommended and will usually be competitive. Because of the importance of user–system interaction in information retrieval, one would be inclined to predict that batch systems will eventually disappear. If an online system is not available, a batch system might be used to produce printed output which could then be used interactively.

Costs of developing an experimental database can be reduced by using, or at least adding to, one already in existence. Additional kinds of indexing, for example, or citations can be added if required by the investigator. However, he or she should ensure that the collection meets the standards of randomness previously described. Also, investigate the possibility that an existing machine collection can be reformated so that it can be processed by an information retrieval package on the investigator's local computer system.

There is a problem of a more general nature with the use of existing databases. If information science is to become a cohesive discipline, knowledge, as in other sciences, must be cumulated on the basis of independent experiments. One cannot confirm or contradict another's general finding by using the same database. There is a grave danger that findings in information retrieval will be the result of idiosyncracies of popular test collections, no matter how well or randomly selected. Confirmation and rejection of conclusions must be based on independent random samples.

Commercial databases such as INSPEC, *Chemical Abstracts Condensates*, *Science Citation Index* can be used in two ways. Either tapes may be purchased and used in conjunction with software developed or purchased for the local computer or the commercial online systems—ORBIT, DIALOG, BRS, etc.—may be used directly. Purchase of tapes is expensive, so that one is usually restricted to a small subset such as a single year, although some database producers have reduced rates for experimental use. In using commercial systems directly, one has their software available and pays only for the time spent searching the system. If the objectives of the experiments can be achieved by using commercial online systems, there seem good economic reasons for choosing this alternative.

5.5 Decision 5: Where to get the queries?

Queries are verbalized information needs, and hence query decisions are really people decisions. This question resolves into three:

(1) What is the source of the original query statement?
(2) Who controls the search process?
(3) Who evaluates the results?

Possible answers to any of the three are:

(1) An actual user of some operational system.
(2) The investigator.
(3) System personnel (operational or experimental).
(4) Any combination of the above.

Clearly, the investigator should not do all three. Such a procedure raises

too many questions about the objectivity of the results. In fact, the investigator should probably function only as a planner. In other words, he or she should not select, search, or evaluate queries, but rather make decisions about procedures for selection, searching, and evaluation.

Great variability has been exhibited by information retrieval experimenters in their method of obtaining test queries. Much the same dilemma arises here as with selection of a database. One may either solicit the co-operation of the actual users of a system or use queries which are in some sense artificial but under greater control of the investigator. The dichotomy is really more of a continuum, where, at one end there is the user-dominated query and search process in which all decisions relating to the initial topics, direction and length of the search, and evaluation of output are controlled by the user, and the experimenter-dominated search, where they are made by the experimenter.

Most experiments lie somewhere between. In Cranfield 2, the authors of source documents framed questions based on their papers and then evaluated the relevance of all references in the original paper. This method, at least, gives an initial relevant set. Other documents in the collection were assessed by judges, not the user. In other experiments, written queries from the history files of an operational system were used with no attempt made to contact the originators. Or queries can be manufactured by artificial means, such as using the title for the query and references for relevant documents.

A problem in using bonafide users is to secure their co-operation, particularly if it means there will be constraints on the search process such as size of file searched or length of search time, and if users are expected to return evaluated output. The user will generally feel that he or she should be receiving something useful in return for his/her time. Free searches on commercial systems with large databases is one inducement. However, some bias will result from this approach, particularly with respect to cost effectiveness. In an environment where users normally pay to do searches, for example, they may be tempted to do for free broader searchers than normal. Some effort to control this factor, such as limiting free offline printing to some maximum number, is probably necessary. With small experimental files, where no large immediate benefit accrues to the user, the most effective approach may be a payment. This method is more successful with indigent students than highly-paid professionals, given the rate most information scientists can afford. Payment of participants should, wherever possible, be included in research grant applications.

Getting assessments of document relevance is an even greater problem than getting queries. With real users, it is best to obtain these immediately after the search and before the user has escaped the premises. If this is not possible, one can again offer an inducement, such as payment or copies of documents to users who complete their evaluations.

Ideally, users should be randomly selected from a pool by the investigator. In practice this is rarely possible. Users are normally self-selected because of the degree of co-operation required of them. The best the investigator can do is to attempt to get a reasonable mix with regard to user traits such as subject background, experience in using the system, and professional level. If random selection is possible then, of course, it should be used. As with databases, conclusions must be restricted to the population from which the

user group comes. If all users in the test are scientists, results cannot be extrapolated to laymen using the same collection, for example.

Sometimes the expertise of searchers or the degree of delegation of the search are independent variables in the experiment. If not, they should either be kept constant for all searches (best approach if the query set is small) or varied randomly (if the query set is large). If search experience is to be held constant, it should be at a high level. The variability that can result from inexperienced searchers may be much greater than the variability resulting from the different treatments under test. Again, this means obtaining the co-operation of experienced systems personnel well before the experiment begins and/or offering payment or other inducement.

Also, it is known that the degree of subject competence affects relevance judgements, i.e. a judge who is familiar with a subject is less likely to accept a document as relevant than one who is not. It is a good idea to get some measure of inter-judge consistency in relevance assessments if the real users are not doing the assessments. Measures similar to those proposed for inter-indexer consistency may be used.

Queries must be clearly stated. If users are the source, they should provide an initial statement of the query in written or taped form. Of course, this may be modified during search strategy construction and/or interactive searching, but the starting point should be clear. If selections are being made from a repository of queries, those that are unclear on any count should be rejected.

5.6 Decision 6: How to process queries?

In comparative searching, it is essential that all things other than the variables under test should be equal. This principle is easier to enforce in a laboratory situation than an operational one. In any test, searchers should be provided with sets of instructions, either as a printed manual or online tutorial. In addition, training and practice sessions for all searchers should be held prior to the experiment. Frequently, problems which would have arisen during the experiment can be spotted at this time. Decide before the experiment what output format is needed and instruct all searchers to this effect.

Unless their use is to be manipulated experimentally, all searchers should have equal access to such search aids as lists of computer commands, index language dictionaries and thesauri, and sample searches. Sometimes, particularly in laboratory experiments, the investigator may wish to make searches of the same query in different languages or systems as alike as possible. Various methods of achieving this control have been used: putting queries into an intermediate language, restricting the search as to time, number of retrieved documents or number of relevant documents, use of a common threshold level with ranked document output.

Many extraneous sources of variation which can occur during computer searching can be eliminated by careful prechecking of the search environment. Are all terminals, data sets, printers, etc., in good operating condition? Are all necessary supplies—paper, pencils, manuals—equally accessible to all search personnel? Will someone, preferably the experimenter, be on the scene to handle the inevitable problems and breakdowns which occur?

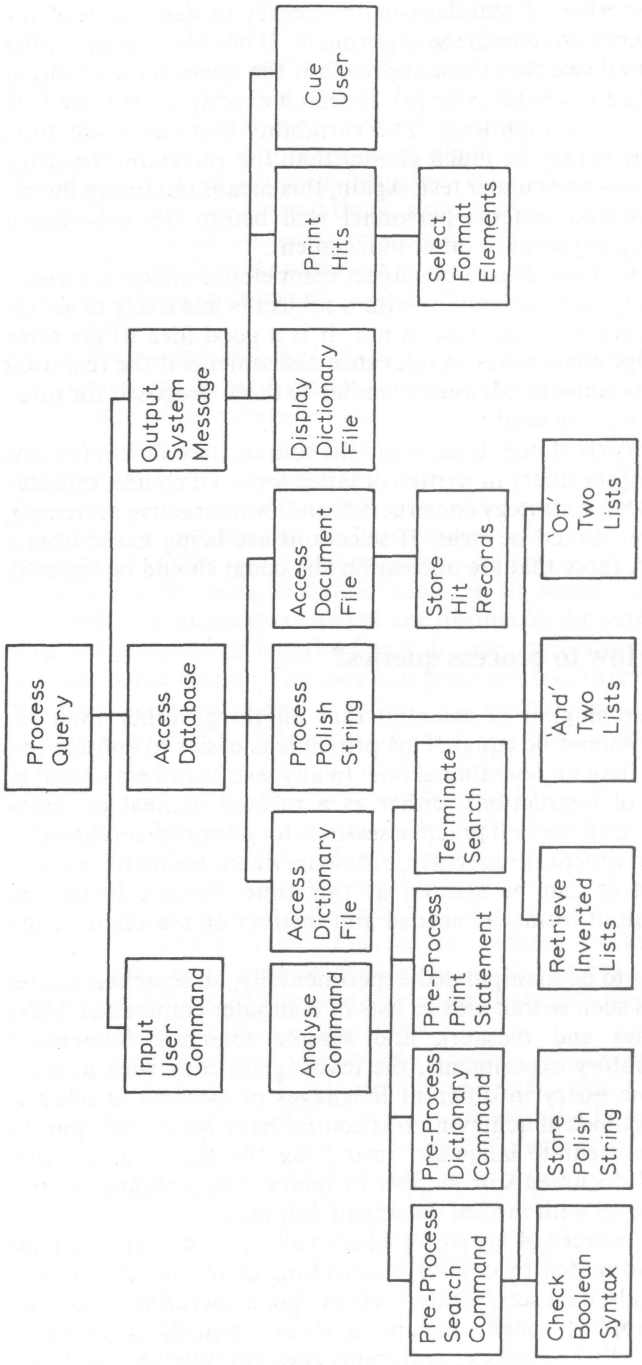


Figure 5.1. Hierarchy chart illustrating top-down analysis for an algorithm to process a boolean statement against an inverted file

During the actual experiment, the investigator must tread a fine line between non-interference and rescue operations.

Although retrieval systems vary greatly in the facilities they provide for searching, a few general comments can be made on ways to improve efficiency or reduce the cost of developing, maintaining, and adapting them.

(1) Do not develop your own retrieval software unless absolutely necessary. There are many retrieval systems now available commercially such as IBM's STAIRS or Stanford's SPIRES/BALLOTS. Consider the possibility of writing pre- and post-processing programs which will permit the searches to be processed by existing packages.
(2) If local software must be developed, employ an experienced computer specialist, at least as a consultant. Insist on a professional product, i.e. software which is:

well-documented
structured
completely debugged before the experiment.

Documentation is essential to ensure that anyone, not just the original designer, can use, maintain, and modify the software. Programs written in a higher-level language such as COBOL should be to industry standards, to provide for portability.

Structured programming implies:

top-down development
modularity
use of standard control structures.

These will be explained in turn.

Top-Down Development: The program is developed as a hierarchy of processes or functions, beginning at the most general level and resolving each process into more specific processes at the next lower level. For example, searching a boolean statement against a database with an inverted file structure could be analysed by the hierarchy chart shown in *Figure 5.1.*

Modularity: Each module consists of about a page of code and corresponds to a single function or process box in the hierarchy chart. Lower level modules must be invoked by modules immediately above them in the hierarchy.

Use of standard control structures: Structured programs are built from three types of building blocks or control structures:

sequence—one step after another
if then else—branching or transfer of control
do while or do until—looping.

The aim of structured programming is to produce programs which are understandable and can be easily debugged, modified, adapted, used in part, etc., by people other than the initial programmer. Given the high turnover rate in computer personnel, this kind of insurance is essential in information retrieval experiments. In addition, one would hope that, if it is necessary to develop an information retrieval system, it can be used in more than one experiment. Hence the need for comprehensible software.
(3) Online retrieval software should provide searchers with the option of at

least two command language modes: learner mode and experienced or abbreviated mode. A 'help' feature, which permits users to access online explanations of the various commands in the retrieval system, is very useful.
(4) Retrieval systems should provide facilities for automatic collection of data needed by the experimenter, for example, number of search statements entered, number of documents retrieved by a search statement, number of postings for any term, search time (both connect time and CPU time).
(5) Retrieval systems to be used in experiments should provide a variety of outputs—title, full citation, abstracts, term or term combination causing retrieval and so on. Facilities should be provided for offline printing of large output sets. If output is to be evaluated, the system should provide it in a suitable form. Users may want to retain the output and so it should be duplicated, one set for the user, one for the experimenter.

5.7 Decision 7: How will treatments be assigned to experimental units?

A complete information retrieval experiment is concerned with assessing the effects of one or more classes of treatments or factors on one or more criterion measures, where the criterion measure is determined for each of a sample of experimental units. For example, the treatments might be different degrees of vocabulary control, the experimental units searches of queries on a database, and the criterion measures recall and precision. Or the treatments might be degree of search delegation to an intermediary, the experimental units online searches of queries on several systems, and the criterion measure total search time. In a partial test, the treatments might be represented by levels of indexer training, the experimental units documents to be indexed, and the criterion measure indexing time. In multi-factor experiments, there is more than a single set of treatments or factors. For example, in complete retrieval experiments, frequently the type of indexing language and the searcher are varied over the query set, giving a two-factor experiment. Or in an online experiment, three factors might be degree of delegation, online system, and searcher. A source of experimental units which is not expected to interact with the factors is called a block. In information retrieval experiments, sources of queries or users such as different libraries might be considered blocks.

In a completely crossed factorial experiment, at least one experimental unit is assigned to each possible combination of factors. Thus, in the two-factor design, indexing language by searcher, in a completely crossed experiment, unique queries would be assigned to each combination of searcher and index language. Thus, if we let g1, g2, g3 represent the three languages, s1, s2, s3, the three searchers, and Q1, Q2, . . . , Q12 twelve query sets, a completely-crossed design would be represented as follows:

	g1	g2	g3
s1	Q1	Q2	Q3
s2	Q4	Q5	Q6
s3	Q7	Q8	Q9
s4	Q10	Q11	Q12

The problem with this approach is that there is such a large variation among queries with respect to recall, precision, and other criterion measures of interest to experimenters that these variations may mask variations caused by the indexing language, which the experiment is supposed to determine.

Another approach is to use a design with repeated measures. As the name implies, this means that the same experimental unit is subjected to the treatments of interest, i.e. each query is searched using all three indexing languages. Such designs permit control over individual differences. Thus, using the same notation as in the previous example, a two-factor experiment (language by searcher) with repeated measures would look like this:

	g1	g2	g3
s1	Q1	Q1	Q1
s2	Q2	Q2	Q2
s3	Q3	Q3	Q3
s4	Q4	Q4	Q4

where Q1, Q2, Q3, Q4 are sets of $n/4$ queries, n being the total number of queries available in the experiment.

If instead of assigning different query sets to each searcher one assigns the same set, then the query has in effect become a third factor in a language by searcher by query experiment.

Repeated measures designs have the advantage that fewer queries are needed for the same reliability. However, they have the drawback of introducing possible 'sequence' effects—the effects of practice, training, learning from a search in one indexing language to a search of the same query in another. In his standard text on experimental design, Winer[18] says:

> 'In experiments where sequence effects are likely to be marked, a repeated measure design should be avoided. In cases where sequence effects are likely to be small relative to treatment effects, repeated measure designs can be used. Randomizing the order of administration tends to prevent confounding of treatment and sequence effects.'

The experimenter must himself judge the magnitude of sequence effects on searchers. One would expect them to be greater with novice than with experienced personnel.

Another way to control sequence effects is by using a Latin square design. A Latin square is an n by n table or array in which the entries in the table are n distinct symbols, assigned so that each appears once in each row and in each column. For example, here are two different 3 by 3 Latin squares:

1	2	3	1	3	2
2	3	1	3	2	1
3	1	2	2	1	3

In experimental design, the rows and columns represent levels of two factors (for example, indexing language and search order). The entries in the body of the table represent experimental units or sets of randomly assembled experimental units (for example, sets of queries). Note that for a Latin square to be used as an experimental design one must have

$$mN(R) = mN(C) = N(Q)$$

where $N(R)$ and $N(C)$ are the numbers of levels in the row and column treatments, respectively, $N(Q)$ is the number of queries, and m is a whole number. If the columns in one of the above two Latin squares represent search order, the rows searchers, and the table entries queries then one can see that the square ensures no two languages will be searched in the same query order. Tables of Latin squares of order up to 12 by 12 may be found in Fisher and Yates[19].

A Greco–Latin square is obtained by combining two Latin squares in such a way (called orthogonally) that each treatment combination occurs the same number of times. This is sometimes useful if there are two factors (for example, indexing language and searcher) as well as a sequence effect. An example of a Greco–Latin square, in which rows represent searchers and columns search order, is:

	1	2	3
s1	Q1g1	Q2g2	Q3g3
s2	Q2g3	Q4g1	Q1g2
s3	Q3g2	Q1g3	Q2g1

Such designs are useful but not always available if the dimension is greater than 5.

Although Latin and Greco–Latin squares ensure that each treatment occurs at each position in the order of administration, it does not totally control sequence effects in the sense that all possible $m!$ sequences of m treatments are observed. This could be achieved by introducing sequence as a fully-fledged factor in a three-factor experiment.

For example, with three indexing languages, there are six possible orderings in which searches can be carried out. Thus, to maintain the Latin square design, one needs six rather than three searchers and six rather than three query sets. The design might be as follows:

	order					
	123	132	213	231	312	321
s1	Q1	Q2	Q3	Q4	Q5	Q6
s2	Q2	Q1	Q4	Q5	Q6	Q3
s3	Q3	Q4	Q6	Q1	Q2	Q5
s4	Q4	Q3	Q5	Q6	Q1	Q2
s5	Q5	Q6	Q1	Q2	Q3	Q4
s6	Q6	Q5	Q2	Q3	Q4	Q1

This design does not control effects due to query sequence. In this and previous designs, random sequencing with respect to queries should be carried out by the searchers. This factor, too, can be formally controlled, but the price would be high in terms of additional complexity of the design and resulting sample size.

Latin squares can be repeated as many times as needed. For example, if the pool of queries was 72 rather than 18, then one might use four Latin squares rather than a single one. Additional squares may be the same as the initial one, others randomly selected, or others obtained by permuting the columns of the first square (balanced squares).

To analyse a Latin square in order to test for differences in the criterion

measure among treatments, one must be able to assume that all interaction effects are negligible. If rows represent searchers and columns languages, there may be an interaction between searcher and language resulting from the fact that certain searchers may find certain languages particularly sympathetic or difficult. In this case, a Latin square should not be used. However, when Latin squares are repeated as part of a larger design, interactions may in part be tested.

In the following design, three balanced Latin squares are used. Search order is indicated by the symbols o1, o2, and o3.

		s1	s2	s3
o1	g1	Q1	Q2	Q3
	g2	Q2	Q3	Q1
	g3	Q3	Q1	Q2
o2	g1	Q2	Q3	Q1
	g2	Q3	Q1	Q2
	g3	Q1	Q2	Q3
o3	g1	Q3	Q1	Q2
	g2	Q1	Q2	Q3
	g3	Q2	Q3	Q1

In this design, it is assumed there are no interactions with the order factor. However, other interactions such as searcher can be tested.

Another example of a repeated Latin square design is given by Keen and Wheatley[3]. Here (see *Figure 5.2*) an incomplete block design is used, in which each block is a Latin square: searcher by order by language. The blocks are incomplete because only a subset of the queries and of the searches occur in each block.

Another kind of sequence effect is involved in the fatigue factor, which can occur in either indexing or searching. Randomizing or otherwise controlling the order in which treatments are applied within a specified time period, for example a day, will reduce this problem. The point at which fatigue sets in can sometimes be determined during preliminary practice sessions. During the actual experiments, scheduling should then terminate activities at this point.

The number of queries is another aspect of experimental design. The more factors under experimental control, the larger must be the query set. Some replication is desirable within each cell (combination of factors). Thus, the more factors to be studied and/or controlled, the larger the sample size required.

For classical, *F*-test, analysis of variance (to be discussed in the next section), Winer[18] provides a method of determining the sample size per cell which will detect a stated minimum difference d among k treatment means at a specified significance level, α and power p. For example, for $\alpha = 0.05$, $p = 0.9$, $d = s/4$, where s is the sample standard deviation obtained from a previous sample, and $k = 5$, the sample size per cell is approximately 56. If d is doubled to $s/2$, the sample size is approximately 14, i.e. it is reduced by a factor of 4. In general, to double the discrimination power of a test one needs to quadruple the sample size.

		Day 1					Day 2				
Order		1	2	3	4	5	6	7	8	9	10
Position		a b	a b	a b	a b	a b	a b	a b	a b	a b	a b
Requests		1 13	2 15	4 14	5 11	3 12	10 17	6 20	8 16	7 18	9 19
Searchers	1	A	B	D	E	C	A	C	B	E	D
	2	B	C	E	A	D	B	D	C	A	E
	3	C	D	A	B	E	C	E	D	B	A
	4	D	E	B	C	A	D	A	E	C	B
	5	E	A	C	D	B	E	B	A	D	C
Requests		20 27	16 30	18 26	17 28	19 29	11 23	12 25	11 24	15 21	13 22
Searchers	6	A	C	B	E	D	A	B	D	E	C
	7	B	D	C	A	E	B	C	E	A	D
	8	C	E	D	B	A	C	D	A	B	E
	9	D	A	E	C	B	D	E	B	C	A
	10	E	B	A	D	C	E	A	C	D	B
Requests		21 33	22 35	24 34	25 31	23 32	30 37	26 40	28 36	27 38	29 39
Searchers	11	A	B	D	E	C	A	C	B	E	D
	12	B	C	E	A	D	B	D	C	A	E
	13	C	D	A	B	E	C	E	D	B	A
	14	D	E	B	C	A	D	A	E	C	B
	15	E	A	C	D	B	E	B	A	D	C
Requests		40 47	36 50	38 46	37 48	39 49	31 43	32 45	34 44	35 41	33 42
Searchers	16	A	C	B	E	D	A	B	D	E	C
	17	B	D	C	A	E	B	C	E	A	D
	18	C	E	D	B	A	C	D	A	B	E
	19	D	A	E	C	B	D	E	B	C	A
	20	E	B	A	D	C	E	A	C	D	B
Requests		41 3	42 5	44 4	45 1	43 2	50 7	46 10	48 6	47 8	49 9
Searchers	21	A	B	D	E	C	A	C	B	E	D
	22	B	C	E	A	D	B	D	C	A	E
	23	C	D	A	B	E	C	E	D	B	A
	24	D	E	B	C	A	D	A	E	C	B
	25	E	A	C	D	B	E	B	A	D	C

Figure 5.2. Incomplete block experimental design used in EPSILON test (from Keen and Wheatley). Indexes (A–E), blocks (1–10) are Latin squares, pairs of blocks (1/2, 3/4, etc.) are balanced

The number of queries in previous information retrieval tests seems to vary from 15 to 300, with values in the range 50 to 100 being most common. Of course, to assess these numbers, one needs to know if queries are completely or incompletely crossed with other factors.

5.8 Decision 8: How to collect the data?

At each stage of a partial or complete information retrieval test, information about various aspects of the experimental process becomes available. What information is actually collected should depend almost entirely on the purpose of the experiment. Extraneous information should not be collected just because it is there. However, if unusual things seem to be happening with some aspect not originally intended as part of the investigation, the investigator may want to collect data as a pilot study for a later full scale investigation.

Data to be collected from information retrieval experiments divides into four categories:

(1) Data about the database—overall characteristics such as size, distribution of indexing term postings, distribution of number of terms per document, distribution in terms of medium, form, source, age, and subject.
(2) Data about people—users, indexers, searchers, authors, managers, etc.—sociodemographic characteristics, subject competence, experience, preferences, values, attitudes.
(3) Data about processes—indexing, searching, using documents, query negotiation—time, cost, number of steps, types of activities and interactions (people–system, people–people).
(4) Data about results—recall, precision, user satisfaction, efficiency, etc.

Data about computerized files can be obtained by appropriate statistical processing. For manual files, the corresponding values may have to be estimated from samples. It is surprising how many operational systems, for example in libraries, keep virtually no statistics on collection size or distribution into different categories.

Computer output from an analysis of the database should be in a form appropriate for incorporation in a report of the study. Clear print, capable of being reproduced, upper and lower case symbols, and some graphics capability should be obtainable from present-day computer installations. Graphics are important because often trends or patterns can be more readily detected in graphical rather than numerical data.

Data on people involved in a study can be collected by observation, using a person or a recording device such as a camera or tape recorder, by interview, either in person or telephone, or by questionnaire. In all cases, the instruments used to record the data should be designed well in advance, preprinted where appropriate, and pretested. The analysis of such data should also be planned in advance, so that the forms can be designed and coded in a manner that will expedite the analysis. This advice is particularly important if analysis is by computer. For example, the investigator should know if the analysis programs can manipulate alphanumeric as well as numeric data (some SPSS implementations, for example, cannot). If alphanumeric cannot be handled, then response categories such as 'excellent, good, average, fair, poor' should be coded with numbers, not letters.

Data from observation and interview records can be keyed in much more rapidly if they are always entered in the same position on the recording instrument. For example, the right-hand side of a questionnaire can contain boxes showing question number–response number pairs. Remember that

there are very few programs which can analyse natural language responses. As far as possible, categorize and code responses.

A good discussion of the advantages, disadvantages and problems of the various techniques for collecting data about people will be found in Kerlinger[20]. Biases which are to be avoided include those caused by the observer's, interviewer's or subject's prejudices, inattention, and misunderstanding; and those related to the Hawthorne effect, i.e. the tendency of subjects under study to perform or respond in a manner different from normal.

Response rate is a problem with mail questionnaires. Some tricks which seem to help are: including a 'reward' with the questionnaire—pencils, notepads, lottery tickets, etc.—follow-up inquiries, particularly by telephone, a description of the purpose and sponsors of the research, promise of a summary of results. This last technique is especially useful with respondents in the same or allied fields.

Some of the methods mentioned above can also be used to obtain a record of a searching or indexing process. Observation by a person is limited by what he or she can see or hear and, at the same time, record. Automatic recording by camera or by tape recorder, for such aspects of searching as query negotiation, is more efficient and reliable. In all cases, there are possible Hawthorne effects, unless the people involved are not informed they are under observation. However, in many institutions this last approach would be considered a breach of privacy. The norm in present-day research practice seems to be to make a visual or audio tape only if the subjects have given their permission.

A record of an online search is obtained automatically from the search printout, which, in an experiment, should always be saved. It is also possible to dump this record onto a disk file for later printing or even automatic analysis. For other processes, subjects can be asked to keep a log or a diary, but this method is less reliable than the printout. Detailed instructions to all subjects can minimize the inconsistencies. Keen and Wheatley[3] have described a useful form of 'index marking' used in the EPSILON tests of printed index searching.

The most intensive data collection usually occurs at the evaluation stage. Forms design and the coding of responses is important here too, if data is to be keyed into a machine-readable file. Mention has already been made of the desirability of supplying users with two output records, one for his/her own use and one to be returned with an evaluation. More general questions about user satisfaction and/or attitude can usually best be handled by questionnaire or interview.

The investigator should look into the possibility of using machine readable instruments for data collection, such as Mark sense cards or optical character recognition (OCR) cards. Although these methods usually have a small error rate associated with them, this may be tolerable in view of the elimination of the input-keying stage. A cost comparison should be made.

Group as well as individual assessments of system effectiveness should be considered. Standard techniques are:

(1) A tape-recorded panel discussion by users, searchers, indexers, and others involved in the experiment.

(2) The Delphi technique, in which individuals are shown an analysis of responses from all members of the group and permitted to revise their own responses. The process is iterated until convergence (agreement) among group members is achieved. No report has been received of a Delphi process which did not converge (for obvious reasons!).

5.9 Decision 9: How to analyse the data?

Analysis of results is either descriptive or inferential. That is, one may simply summarize the data obtained or one may generalize and make predictions from it about larger sets of data or populations.

As mentioned earlier, the techniques of statistical inference and decision-making are based on the assumption that the data constitutes a random sample from the population, i.e. a sample selected in such a way that each possible sample of the same size has the same probability of occurring. In practice, we cannot always guarantee that this condition has been met. A sample is usually considered suitably random if some kind of chance mechanism has been used in its selection and there are no apparent biases.

It is only in the past few years that inferential rather than descriptive methods have been used at all widely in information retrieval testing. One reason for earlier neglect may have been that information scientists were not familiar with statistical inference. Another is that sample document and query sets were distinctly non-random. However, the importance of randomization and experimental design is increasingly recognized in retrieval experiments and so inferential tests should be more prevalent in the future. The value of statistical inference lies in its generalizing potential. Unless information science is able to derive general results or 'laws', it will remain a very primitive science.

Descriptive methods

Descriptive methods encompass:

(1) The various graphical and tabular displays of variable frequencies and relationships, such as the recall–precision curve, which have long been part of information retrieval test methodology.
(2) The calculation of descriptive statistics measuring central tendency, variability, association, and other characteristics.

Measures of central tendency include:

the arithmetic mean, or average value;
the median, or middle value;
the mode, or most frequent value.

Measures of variability include:

the variance, or averaged squared distance of the observations from their mean;
the standard deviation, or square root of the variance;
the range, or difference between the smallest and largest values;

the interquartile range, or range within which the middle 75 per cent of the observations lie.

Measures of association will be considered later.

The appropriate measure to use depends to a large degree upon the scale of the observations. Four scales of measurements are distinguished in social science research:

nominal—names or categories;
ordinal—ranks;
interval—numbers;
ratio—numbers with a zero point.

The last two are the true quantitative scales. In statistical analysis, the distinction between interval and ratio is of no particular value. A much more important distinction so far as type of analysis is concerned is between discrete and continuous variables (i.e. variables which are counts versus variables which can take any real value in an interval).

Arithmetic operations can be properly applied only to numbers. Thus, means and standard deviations should not be calculated for ordinal data, but medians and interquartile ranges, which require only a ranking of the observations, may be.

Any variable that is essentially a count—such as number of relevant documents—or some function of counts—such as recall and precision—can be considered a ratio scale. No value judgement is implied by saying that one method has twice the precision of a second, one is simply stating a numerical fact about the ratio of the two values. It does not necessarily mean that the first method is twice as good as the second, any more than a height of 8 feet is twice as good as a height of 4 feet. Appropriate methods depend on the scale of the observations, not their value to the user or other individual. This point is important because many information retrieval investigators have shied away from classical statistics when there was no real reason to do so. Any set of numbers—counts, proportions, logarithms—can be averaged. Normality is not essential. It does not affect the validity of descriptive statistics, although it may affect their value. Normality is important in determining appropriate tests in statistical inference.

There are, however, problems with averaging recall and precision over a set of queries. These relate to the method of averaging. Two kinds are possible:

average of numbers (microaveraging);
average of ratios (macroaveraging).

If four queries have the precision values shown in *Table* 5.1,

TABLE 5.1

Query No.	*No. of retrieved references*	*No. of relevant references*	*Precision*
1	25	10	0.6
2	5	2	0.4
3	10	5	0.5
4	1	1	1.0
Total	41	18	2.5

Microaverage of precision is $18/41 = 0.439$. Macroaverage of precision is $2.5/4 = 0.625$.

The choice of averaging method hinges on whether one wishes to give documents or queries equal weight in the averaging process. However, if the averages are to be used as sample estimates of population values, as discussed in the next section, then the microaverages should be used, as these have the statistically desirable property of maximum likelihood (see Tague and Farradane[15]). Another advantage of microaveraging is that one does not usually have to deal with the undefined value 0/0. In macroaveraging, one can either set such ratios equal to 1 or throw out the query. Neither course is really satisfactory.

Another problem, thoroughly discussed by Sparck Jones[21] and others, relates to the recall–precision graph. Given ordered document output for a set of queries, the recall–precision graph will depend on both the measure of document–query similarity (the scores) and the choice of points to be displayed on the graph. As described in Section 5.3, there are a number of ways in which the document query similarity can be measured. These include:

(1) Co-ordination level, i.e. the number of terms matching between query and document.
(2) Cosine coefficient and other weighting functions.

Documents may be ranked on the basis of any of these measures.

In order to construct a recall–precision graph, the points at which recall and precision values will be averaged over queries and displayed on the graph must then be determined. There are four possibilities:

(1) Average recall and precision across queries at fixed document–query similarity scores. This method works well with co-ordination level scores but creates problems with document–query weights which assume a large number of values.
(2) Average recall and precision across queries at fixed document ranks. This method is useful when the document–query scores assume a large number of values.
(3) Average recall and precision values at either fixed scores or fixed ranks and then interpolate precision at standard recall values, for example 0,0.1,0.2, . . . , 0.9,1. This gives a smoother curve than Methods 1 and 2. Two interpolation methods have been suggested:
    (a) linear interpolation,
    (b) interpolation to the left between averaged recall values ('pessimistic' interpolation).
(4) Interpolate precision values at standard recall values for each query and then average precision values over the queries.

When the number of terms matching between document and query (co-ordination level) is an independent variable, a set of average recall and precision values can be obtained for a query at each degree of match, i.e. at 1,2,3, . . . matching terms. A problem arises because not all queries have the same number of terms, so that the average will be over different numbers of queries at some co-ordination levels. One can examine only subsets consisting

of queries with the same number of terms, but this makes difficult an overall assessment of performance. Co-ordination level averaging is best used when there is not too much variation in number of terms from query to query.

Similarly, for query–document weight cut-off or document rank cut-off, the set of scores or ranks may differ from query to query, so that these methods are best used when there is not much variation in the range of scores or size of output from query to query.

If precision at standard recall values is used, then selection, interpolation, and extropolation may be needed to obtain a single precision value for each recall value. The two possibilities for interpolation and extrapolation are:

linear interpolation/extrapolation
'pessimistic' interpolation/extrapolation, i.e. use the precision value for the next higher recall point.

The differences among the various methods are illustrated in a simplified example, *Figure 5.3*, in which a recall–precision curve is calculated using four different methods:

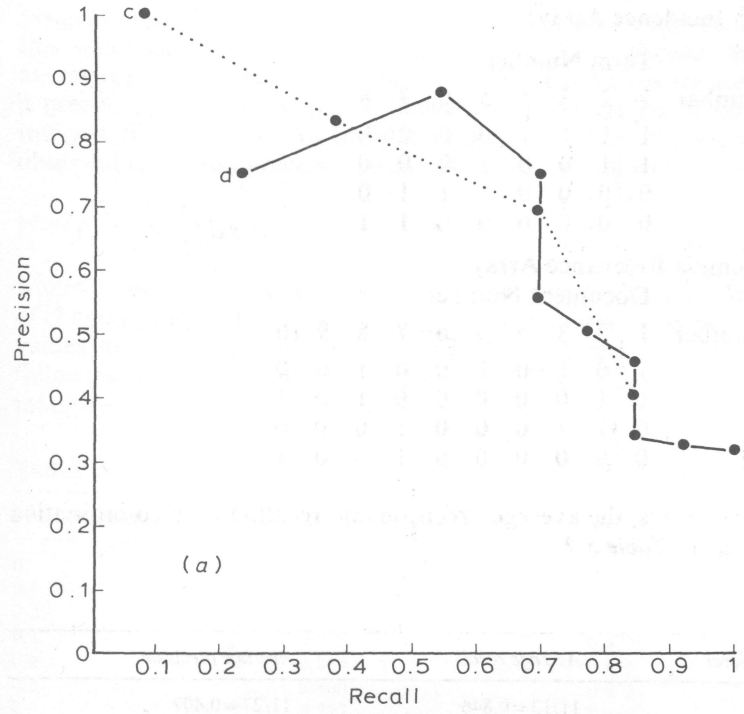
(1) Retrieval cut-off by document–query scores, with co-ordination level scoring and microaveraging of recall and precision.
(2) Retrieval cut-off by document rank, with cosine coefficient scoring and microaveraging of precision and recall.
(3) Interpolation of average precision values at standard recall values, from the data in 2, using linear interpolation.
(4) Interpolation of average precision values at standard recall values, from the data in 2, using pessimistic interpolation.

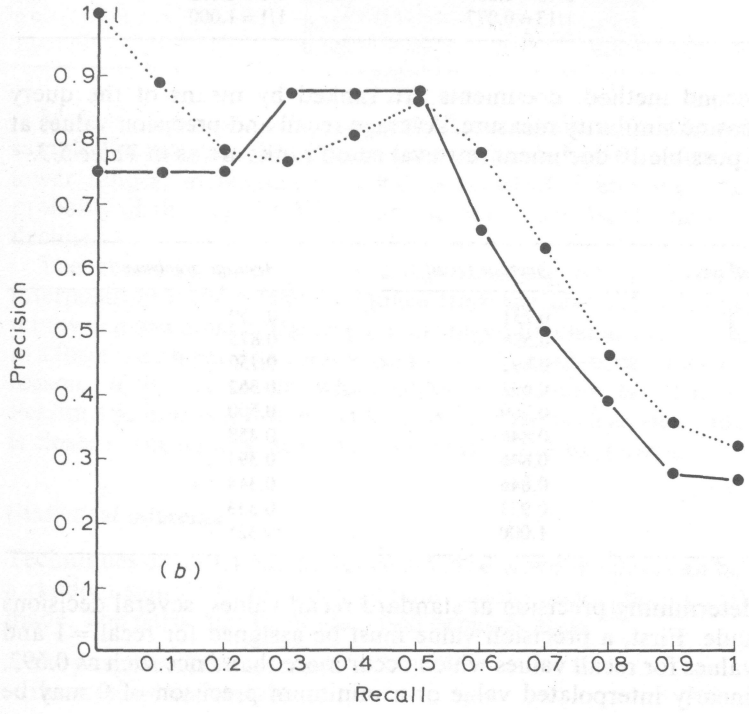
The data which generated the curves in *Figure 5.3* are given by the following arrays:

Document–Term Incidence Array:

	Term Number							
Document Number	1	2	3	4	5	6	7	8
1	1	1	1	1	0	0	0	0
2	1	1	0	0	1	1	0	0
3	0	0	0	0	0	1	1	1
4	0	0	0	1	0	1	1	0
5	0	1	1	0	0	0	0	0
6	1	0	0	0	0	0	0	1
7	0	0	0	0	1	1	1	1
8	0	1	1	1	0	0	0	0
9	0	1	0	0	0	1	1	0
10	1	0	0	0	1	0	0	0

Figure 5.3 (*opposite*). Recall–precision curves for a simplified retrieval output, using four methods of determining the points at which recall and precision will be averaged.
(*a*) c . . . . . Co-ordination level points;
d——— Document rank points. (*b*) l . . . . . Standard recall points, linear interpolation; p——— Standard recall points, pessimistic interpolation


c
d
Precision
Recall
(a)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


l
p
Precision
Recall
(b)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Query–Term Incidence Array:

	Term Number							
Query Number	1	2	3	4	5	6	7	8
1	1	1	1	1	0	0	0	0
2	1	1	0	0	1	0	0	0
3	0	0	0	0	1	1	1	0
4	0	0	0	0	0	0	1	1

Query–Document Relevance Array:

	Document Number									
Query Number	1	2	3	4	5	6	7	8	9	10
1	1	0	1	0	1	0	0	1	0	0
2	1	1	0	0	0	0	0	1	0	1
3	0	1	1	0	0	0	1	0	0	0
4	0	0	0	0	0	0	1	0	0	1

From these arrays, the average precision and recall at each co-ordination level will be as in *Table 5.2.*

TABLE 5.2

Co-ordination level	*Average recall*	*Average precision*
1	11/13 = 0.846	11/27 = 0.407
2	9/13 = 0.692	9/13 = 0.692
3	5/13 = 0.385	5/6 = 0.833
4	1/13 = 0.077	1/1 = 1.000

In the second method, documents are ranked by means of the query document cosine similarity measure. Average recall and precision values at each of the possible 10 document retrieval cutoff ranks are as in *Table 5.3.*

TABLE 5.3

Document cutoff level	*Average recall*	*Average precision*
1	0.231	0.750
2	0.538	0.875
3	0.692	0.750
4	0.692	0.562
5	0.769	0.500
6	0.846	0.458
7	0.846	0.393
8	0,846	0.344
9	0.923	0.333
10	1.000	0.325

Before determining precision at standard recall values, several decisions must be made. First, a precision value must be assigned for recall = 1 and precision values for recall values which occur more than once such as 0.692. Either a linearly interpolated value or a minimum precision of 0 may be

assigned to recall = 1. Because of the erratic nature of low recall values for this small sample of 4 queries, the latter course is chosen. Similarly, the average of the precision values for repeating recall values are used, obtaining a precision of 0.656 for recall = 0.692 and 0.398 for recall = 0.846. Linear interpolation values p^* for the standard recall values r^* are obtained from observed recall and precision points by using the following formula:

$$p^* = p_1 + \frac{r^* - r_1}{r_2 - r_1}(p_2 - p_1)$$

where r_1 and r_2 are the recall values immediately to the left and to the right of r^* and p_1 and p_2 the corresponding precision values. 'Pessimistic' precision values are those associated with the recorded recall value immediately following (i.e. greater than) the standard one. The values obtained by the two methods are shown in *Table 5.4*.

TABLE 5.4

Standard recall	*Linear precision*	*Pessimistic precision*
0	1.0	0.75
0.1	0.892	0.75
0.2	0.784	0.75
0.3	0.778	0.875
0.4	0.819	0.875
0.5	0.860	0.875
0.6	0.787	0.656
0.7	0.640	0.5
0.8	0.460	0.398
0.9	0.352	0.333
1.0	0.325	0.325

Methods based on document cut-off are particularly vulnerable to small sample fluctuations. The somewhat unusual behaviour of precision in the lower ranges, increasing to 0.860 at recall = 0.5 and then decreasing, is probably of this nature. With large samples, one usually finds a monotonic decline.

The recall–precision curve using standard recall points with linear interpolation is the furthest removed from the actual data, in the sense that it may contain none of the original averaged precision values. However, it is in a form which permits comparison to other recall–precision curves. For this reason, it is preferable when different systems are being compared. Pessimistic interpolation provides a more conservative view, and, because it is closer to the data, a curve which is usually not so smooth.

Statistical inference

Techniques of statistical inference are used when the data can be considered a random sample from which generalizations about the population will be made. The particular technique used depends upon

the purpose of the research;
the scale of the variables.

Most information retrieval experiments are carried out for one or more of the following purposes:

estimation;
comparison;
exploring relationships;
prediction.

To describe all the statistical tests which have been proposed for these problems would require many volumes. This chapter, for the most part, will simply indicate, for each of the four categories above, the factors which determine the tests to use, rather than give the details of the test themselves. These can be found in many standard statistical texts. Good introductory texts are Mendenhall *et al.*[22] and Winkler and Hayes[23]. A rigorous mathematical development will be found in Kendall and Stuart[24]. Noether[25] is useful for non-parametric statistics.

The scale of measurement of the data determines whether classical or non-parametric tests are appropriate. Classical statistical techniques such as T tests, F tests (analysis of variance or ANOVA), regression, and product–moment correlation may be applied if:

(1) the population is known to be normally distributed, or
(2) the data is continuous or discrete with a large set of values and the sample size is large.

The inclusion of this second category is justified by the Central Limit Theorem of statistics, which says that, even though the population is non-normal, the sample means will be approximately normal for large samples. The Chi Square goodness-of-fit test or the more efficient Kolmogorov–Smirnov test may be used to test whether or not a sample appears to come from a normal distribution.

Estimation

In estimation, one uses a sample statistic (the estimator) to estimate a population parameter. By a sample statistic we mean some quantity which is calculated from a set of sample observations. For example, the average precision for all queries put to a retrieval system or the proportion of all users of a system who are satisfied with it. A sample estimator, such as the sample mean, is said to be unbiased if its expected value is equal to the population parameter being estimated.

The estimator is a random variable, i.e. its value will vary from sample to sample and will not necessarily be equal to the population parameter. Thus, in inference, it is useful to associate some kind of probabilities with the errors that may be made in using the estimator in place of the true population value. There are two ways of making such probability statements—standard errors and confidence intervals.

The standard error of an estimator is its standard deviation. It indicates how much the estimator varies from sample to sample. The greater the standard error, the lower the reliability of the estimator. The reliability of an estimator is thus related to its probability distribution. Some useful theorems of probability theory provide information about two of the most important

estimators—the sample mean and the sample proportion—when the sample size is large. It can be shown that, in both cases, these estimators are normally distributed and that the means are equal to the corresponding population values. The standard deviations, which are the standard errors, can be approximated by the following formulas:

standard error of the sample mean: $s/\sqrt{n}$, where s is the sample standard deviation and n is the sample size;

standard error of the sample proportion: $\sqrt{p(1-p)/n}$, where p is the sample proportion and n is the sample size.

The standard error, by itself, does not mean much. It is more usefully employed in setting up confidence intervals. A confidence interval is an interval or range on either side of the sample estimator which contains the population value with a given confidence. Confidence is usually expressed as a percentage between 0 and 100. A 95 per cent confidence interval, for example, is an interval determined in such a manner that 95 times out of 100 it will contain the population value. With large samples, a 95 per cent confidence interval for the population mean or population proportion will be approximately 2 standard errors (more exactly 1.96) on either side of the sample value. A 99 per cent confidence interval will be approximately 3 (more exactly 2.57) standard errors on either side of the estimator. For example, suppose, in a survey of users attitudes to an online retrieval system, it was found that 96 out of the 120 users surveyed were satisfied with the service they received. The standard error is thus given by

$$\left[\frac{96/120(1-96/120)}{120}\right]^{1/2} = 0.036$$

and a 99 per cent confidence interval for the satisfied proportion of the population would be

$$\frac{96}{120} \pm 2.57(0.036) = (0.707, 0.893)$$

These methods assume, also, a large population. Other methods must be used to set up confidence intervals for small populations.

Tague and Farradane[15] have shown that if one estimates system recall and system precision by searching random samples of n queries against m documents and calculates estimated recall, say, by using microaveraging to obtain a sample estimator $\hat{\rho}$ of recall, the estimator will have a standard error of approximately

$$\left(\frac{\rho(1-\rho)}{mn\gamma}\right)^{1/2}$$

where ρ is the average system or population recall, γ is the average system generality, m is the size of the database, and n is the number of queries in the sample. This may be approximated by

$$\left\{\frac{\sum_{i=1}^{n} a_i \sum_{i=1}^{n} c_i}{(\sum_{i=1}^{n} (a_i + c_i))^3}\right\}^{1/2}$$

Similarly, the standard error for the system precision estimator, $\hat{\pi}$, is

$$\left(\frac{\pi(1-\pi)}{mn(\rho\gamma/\pi)}\right)^{1/2}$$

and is approximated by

$$\left\{\frac{\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i}{(\sum_{i=1}^{n} (a_i+b_i))^3}\right\}^{1/2}$$

a_i is the number of relevant and retrieved documents for the ith query, b_i is the number of non-relevant and retrieved documents for the ith query, c_i is the number of relevant and non-retrieved documents for the ith query. The estimators $\hat{\rho}$ and $\hat{\pi}$ will be approximately normal for large samples and hence their standard errors can be used to set up confidence intervals in the manner described above.

Comparison

In statistical inference, comparisons are carried out by hypothesis tests, i.e. tests of the null hypothesis that there is no difference among or between the treatments or factor levels. If the null hypothesis is rejected, then a difference has been shown to exist, with specified probabilities of making a wrong decision. The general procedure for hypothesis testing is as follows:

(1) State the null hypothesis H0 and the alternative hypothesis H1. The null hypothesis is generally the hypothesis of no difference, the alternative hypothesis may specify a difference in either direction or a difference in one direction only (e.g. one value > the other).
(2) Set a significance level, usually denoted α. The significance is the probability the null hypothesis will be rejected when it is actually true. It limits the probability of such Type 1 errors. A Type 2 error occurs when the null hypothesis is accepted when it is false. Its probability is denoted β and $1-\beta$ is called the power of the test. For a fixed sample size, usually as α is increased β decreases. In the usual hypothesis test only α is limited; however β may also be limited, in some tests, by an appropriate sample size. The usual significance levels are 0.05 or 0.01.
(3) Select a random sample from the population or populations being tested.
(4) From the sample values calculate the value of an appropriate test statistic. Like an estimator, the test statistic is a random variable. Its distribution under the null hypothesis must be known. Some commonly occurring test statistics are Z (standard normal deviate), Student's T, F (in ANOVA), and Chi square.
(5) Compare the value of the test statistic with the critical value or values in tables of the appropriate probability distribution under the null hypothesis. The critical value will be that table value which will give a probability of α of rejecting H0.
(6) If the test statistic value lies outside (usually greater than and/or less than) the critical value or values, the null hypothesis is rejected. Otherwise it is accepted.

Hypothesis tests which may be used to compare two factors or treatments—for example, two search strategies, two indexing methods, two kinds of

users—are shown in *Table 5.5*. The appropriate conditions for each test are also indicated.

TABLE 5.5

Design/Variable type	*Test*
Independent samples	
Normal, equal variances	*T* test
Continuous or discrete with many values, large sample (>30)	*Z* test
Continuous, discrete, some ordinal	Wilcoxon–Mann–Whitney test, Median test
Dependent samples	
Normal	*T* test of differences
Continuous or discrete with many values, large sample (>30)	*Z* test of differences
Continuous, discrete, some ordinal	Sign test

By dependent samples, we mean that the samples under the two treatments are matched in some fashion—for example, two indexing languages applied to the same set of documents, two search strategies used with the same set of queries.

The non-parametric tests such as the Wilcoxon–Mann–Whitney test and the Sign test have generally been developed on the basis of the assumption that the data are continuous. Modified procedures have been developed for situations in which the data are discrete and ties are present. Noether[25] points out that, in the long run, the proportion of times that H0 is rejected when true corresponds to the chosen significance level. Many texts also suggest that these tests can be applied to ordinal data. However, because the derivation of the tests depends on an assumption of continuous or discrete data, this approach should not be taken unless it makes sense to consider the ranks as merely representing an underlying continuous scale. For example, one might ask users to rank documents from two search strategies as to relevance and use a Wilcoxon test to compare the results if it was felt that the ranks represented a continuous relevance weight. Whether this assumption is justified is a theoretical rather than pragmatic question.

The condition that the population variances are equal, required for the Student *T* test, may be tested using an *F* test.

As an example of both a classical and a non-parametric test for the same hypothesis, consider the following test comparing two indexing languages. Ten queries are searched using both language A and language B. The null hypothesis and alternative hypothesis are:

H0: $\Pi_1 - \Pi_2 = 0$

H1: $\Pi_1 - \Pi_2 \neq 0$ where π_1 and π_2 represent the average precisions for the two languages. The significance level is set to 0.05. The sample precision values for the ten queries for each method are:

Method A: 0.65, 0.18, 0.32, 0.49, 0.64, 0.30, 0.86, 0.22, 0.35, 0.20
Method B: 0.78, 0.19, 0.33, 0.47, 0.66, 0.77, 0.97, 0.21, 0.36, 0.13

Since the sample size is small, if one were not certain of the normality of

precision, the Sign test could be used here, as samples are dependent. To determine the test statistic K, a + or a − is assigned to each query depending on whether the precision score for B is greater than or is less than that for A. This gives the following sequence of signs:

$$+ + + - + + + - + -$$

The test statistic K is the number of plus signs, so that $K=7$. K has a binomial distribution with parameters $n=10$ and $p=1/2$.

Since the test is two sided, i.e. the null hypothesis will be rejected for both high and low values, a two-sided critical region is needed. With discrete distributions taking few values, it is not always possible to define a critical region which will have a probability exactly equal to the significance level. Here, the binomial distribution provides the following probabilities under H0:

$$P(K<2 \text{ or } K>8) = 0.022$$

$$P(K<3 \text{ or } K>7) = 0.1$$

With a significance level of either 0.022 or 0.1 we would accept the null hypothesis with a K value of 3. Thus, we can conclude that for $\alpha=0.05$ it will also be accepted.

If we are willing to assume a normal distribution for precision scores, perhaps from previous evidence, then the Student T test can be applied. Here, instead of assigning a + or − to each query, we determine the difference between the A and B precisions. This gives the following differences:

$$0.11, 0.01, 0.01, -0.02, 0.02, 0.47, 0.11, -0.01, 0.01, -0.07$$

The test statistic is

$$T = \sqrt{n}\bar{D}/S = 1.325$$

where $\bar{D}$ is the average of the 10 differences, S is their sample standard deviation, and n is the sample size, 10.

The critical region for this test, determined from table of the T distribution, is $T>2.26$ or $T<-2.26$. Thus, the same conclusion as in the Sign test is reached—accept H0.

Though one may be surprised at the lack of a significant difference for this data, it must be remembered that small samples in general require very large differences to attain significance. Essentially, the test is saying that the observed superiority of Method B could arise from random fluctuations among queries. Obviously, the T test is more sensitive to the magnitude of the differences.

Tests for comparing three or more treatments are shown in *Table 5.6*.

ANOVA procedures also exist for many more complicated multifactor designs, such as Latin squares. Until recently, corresponding non-parametric tests did not exist. However, there is active development in this area, and the investigator is advised to consult the recent statistics literature.

If variances of the samples under different treatments appear to be unequal, they may be stabilized by a transformation of the original observations. Some common transformations are the square root, the logarithmic, and the arcsin (see Winer[18] for details). Winer also points out

TABLE 5.6

Design	*Type of variable*	
	Approx. normal, equal variances	*Continuous, discrete, some ordinal*
Single factor		
Independent samples	One-way ANOVA	Kruskal–Wallis test
Dependent samples	One-way ANOVA, repeated measures	Friedman test
Complete blocks	One-way ANOVA, complete blocks	Noether's T_n test
Incomplete blocks	One-way ANOVA, incomplete blocks	Durbin test

that the F test is relatively insensitive to moderate departures from normality. Thus, it may be used when the data are only approximately normal. Many of the variance stabilizing transformations also make the data more normal.

In general, data consisting of counts, e.g. number of relevant documents, or times, e.g. search time, should be analysable by parametric methods. The arcsin transformation is useful in stabilizing the variances and improving the normality of proportions such as recall and precision. Times which are skewed towards low values can have their distributions improved by the logarithmic transformation.

Following a significant ANOVA, i.e. a significant difference in treatments, the experimenter may wish to test which particular treatment pairs differ. A number of tests are available for such contrasts: the Newman–Keuhls, Duncan, Tukey, and Sheffé tests. Details may be found in Winer.

Wherever possible, a parametric test is to be preferred to a non-parametric one because of its great efficiency. Pittman (see Noether) defines efficiency as follows:

> 'If we have two tests of the same hypothesis and significance level and if for the same power with respect to the same alternative one test requires a sample size $N1$ and the other a sample size $N2$, the relative efficiency of the first with respect to the second is given by $e = N2/N1$.'

Noether gives specific examples of the efficiency of non-parametric tests against normal curve alternatives. The asymptotic (i.e. large sample) efficiency of the T_n, Kruskal–Wallis, Durbin, Friedman, and Wilcoxon–Mann–Whitney tests will not fall below 0.864 and may be as high as 0.955. The Sign test, however, has an efficiency of only 0.64.

Another advantage of parametric tests is that they are easier to compute. Most non-parametric tests require ranking the observations, an operation whose time is proportional to n^2, or at least $n \log n$. Parametric tests, on the other hand, are based on adding and squaring—operations whose time is proportional to n. For large samples, this difference may be important.

Exploring relationships

Exploring relationships may involve either:

(1) Determining if two variables are related or independent, e.g. is search time related to searcher experience?
(2) Estimating the degree of relationship between them, e.g. what is the correlation between the frequency of use of a document and its age?

The usual method of determining whether or not two variables are related when at least one of the variables is non-quantitative is by means of the Chi square contingency table test. For example, a sample survey of personnel in an organization gave *Table 5.7.*

TABLE 5.7

Employment class	*Used online retrieval systems*	
	Yes	*No*
Manager	5	28
Scientist/engineer	30	4
Technician	15	29
Clerical	10	19

A Chi square statistic calculated for the table would indicate whether system use was dependent on or independent of employment classification. (In fact, the null hypothesis of no relationship is rejected.)

When neither variable is qualitative, the relationship between two variables can best be expressed by a single number, a measure of association. The hypothesis of no relationship then reduces to a test of the hypothesis that the measure of association is 0.

If both variables are continuous or discrete with many values, the product moment or Pearson correlation coefficient may be used. If sample sizes are large, a transformation of the coefficient will have an approximately normal distribution. This may then be used to test the hypothesis that the correlation, i.e. the linear relationship, between the two variables is 0. It is also possible to test whether or not two samples come from populations with the same correlation, for example, to test whether the correlation between search time and number of retrieved references was the same for two different online systems.

If one or both of the variables are measured on a scale which is ordinal or better, then a rank correlation coefficient, either Kendall's tau or Spearman's rho, may be used to measure association. Hypothesis tests similar to those for the product moment coefficient may be applied. The relative efficiency of the test of no correlation using tau as opposed to the product moment correlation when populations are normal is 0.91.

Prediction

Regression techniques are used to predict the value of a dependent variable from other independent variables. In linear regression, the dependent variable is expressed as a linear function of another variable, for example, cost of a search as a function of number of retrieved documents. In multiple linear regression, it is expressed as a linear function of several other variables, for example search time as a function of number of search statements, number of retrieved documents, and number of unique descriptors. In non-linear regression, it is expressed as a non-linear function of one or more other

variables, for example vocabulary size as a logarithmic function of collection size. Confidence intervals may be set up for predicted values; however, the accuracy and reliability depends upon an assumption of at least approximate normality.

Although superficially like the preceding problem, forecasting future values of some variable on the basis of past values is not really amenable to regression techniques. This is because regression is based on the assumption of independent observations. Time series, such as daily use of a system or monthly recall/precision figures for an SDI profile, are obviously dependent observations—one day's or month's value is related to previous ones. Time series analysis, which consists of analysing a series in terms of trends, periodic or seasonal components, and random fluctuations, is discussed in detail in a number of monographs, for example Gilchrist[26] and Box and Jenkins[27].

Implementation

Finally, there is the question of the medium for data analysis. There are two ways:

(1) Manual tabulations, possibly using hand calculators. This is convenient in the sense that it can be done internally, but may not in the long run be the least expensive method. It is necessary, of course, for the analysis of non-formated data. Manual tabulations have a very high probability of error, so that, to be sure of results, all calculations must be verified. This can be very tedious, particularly if results do not tally the first time.
(2) Computer-based statistical packages. The chance of error is much reduced here, though, of course, data input must still be verified. The best known statistical packages are SPSS (Statistical Package for the Social Sciences), SAS (Statistical Analysis Package), and BMD (Biomedical Computer Programs), and it is probably best to use one of these if you are carrying out a wide range of different types of analysis on the same data. The actual tests available with these packages vary, to some extent, from installation to installation. For example, some installations have non-parametric tests not described in the SPSS Manual. A useful introduction to the three packages listed above will be found in Moore[28].

It is important, however, to understand the function of the different tests in the packages. Their very comprehensiveness makes them susceptible to misuse. Anyone contemplating the use of statistical packages should study the manual carefully prior to data collection. Much time and expense at the data analysis stage can be saved by collecting data in a form that is amenable to entry into an SPSS or other package file. Basically, data is entered case by case, each case consisting of several fields defining characteristics of the case. Sometimes there is a problem in deciding what is a case. For example, in a study of retrieval, is a case a searcher, a user, a query, a search, a search statement. It all depends on the purpose of the analysis. A case should be the simplest, most atomic experimental unit to be examined in the study. If users have several queries and queries consist of a sequence of search statements, and if interest is in the effectiveness of various ways of structuring search statements, then a case is a single search statement.

5.10 Decision 10: How to present results?

Information retrieval experiments should be written up as experiments. This rather obvious recommendation is not always followed in practice. In many reports, one does not realize an experiment has been performed until half way through the paper. The first part is all background.

The various aspects of an experiment are generally described in the following order:

Purpose of the experiment.
Background for the experiment.
Methodology.
Presentation of results.
Summary and conclusions.

The purpose of the experiment should be described both in general and specific terms, i.e. the general problem or hypothesis being investigated and its realization, for this experiment, in terms of operational variables.

The background section should provide justification for the experiment. What previous work has been carried out in this area? Why is the present study needed? What led the investigator to undertake the work? Only references that specifically relate to the problem under study should be included.

Methodology can usually be subdivided into two sections: the test environment and the test procedures. The environment refers to the characteristics of the documents, document surrogates, queries, users, searchers, equipment, etc., used in the test. These should be characterized in detail, as the generality of the results depend on these aspects.

Procedures relate to the actual methods used to select the sample (experimental design), run the experiment, collect the data, and analyse the results. Procedures should be described in sufficient detail that another experimenter can repeat the experiment. However, aspects which have been described previously, in generally accessible documents (i.e. not in private communications), such as search algorithms or statistical tests, should simply be referenced.

In the results section, the investigator attempts to summarize verbally what the experimental results have shown, not just present pages of tables. Detailed computations or mathematical derivations should be relegated to an appendix and their conclusions only incorporated in the text. Similarly, detailed results and analyses, such as a query by query failure analysis, should also be in the appendix.

The final section should serve to review, reiterate, and summarize what has gone before. Remember that this section is all many people read!

Some small but important matters remain to mention. Symbols should not be introduced without precise definition. Even conventional symbols can have several interpretations: for example, π can represent either a mathematical constant, or a product operation, or system precision.

Both the horizontal and vertical axes in graphs should be labelled and the scale indicated. Unless results are really voluminous, graphs should be accompanied by tables showing the specific values used to construct them. All graphs and other figures should have legends. Other investigators may

want to attempt to reproduce the graphs. If they are impressionistic rather than exact, and have no accompanying tables that is not possible.

The traditional method of presenting experiments is not a chronological narrative. It may seem to put the cart before the horse in that the true purpose of the experiment may not have been clearly defined until after some initial 'messing around'. This can be indicated in the 'Background' section, but a reader should not be subject to long personal histories. The important questions to the reader and to the discipline are:

What was the problem?
How was it solved?
What is the solution or conclusion?

In presenting results and conclusions, the experimenter must be careful to avoid exaggerated claims. It is difficult not to have a personal interest in confirming a particular hypothesis. However, this tendency must continually be restrained and objectivity sought, particularly in evaluating results. Nothing should be claimed that could not be verified by an independent investigator. On the other hand, the investigator should not neglect to point out results that are interesting or unusual, though not adequately tested. These frequently provide the seeds for future investigations.

To summarize, the presentation of results must maintain a delicate balance between completeness and conciseness. Previous reports which seem particularly interesting and comprehensible should be studied as models for the presentation of results.

This paper has given some guidelines and practical suggestions for investigators embarking on an information retrieval experiment. Some of the recommendations may be questioned by others in the field. Some are based on the author's personal experiences or the experiences of her students. The model has been prevailing practice among people the author considers to be serious investigators in the social, biological, and physical sciences. Information retrieval experiments must meet the same criteria if information science is to become a respectable area of scientific inquiry.

References

1. CLEVERDON, C. W., MILLS, J. and KEEN, E. M. *Factors Determining the Performance of Indexing Systems*, 2 Vols, College of Aeronautics, Cranfield (1966)
2. BROOKES, B. C. A measure of categorical dispersion, *Information Scientist* **11**, 11–17 (1977)
3. KEEN, E. M. and WHEATLEY, A. *Evaluation of Printed Subject Indexes by Laboratory Investigation*, British Library Research and Development Report 5454, College of Librarianship, Aberystwyth, Wales (1978)
4. SALTON, G., WONG, A. and YU, C. T. Automatic indexing using term discrimination and term precision measurements, *Information Processing and Management* **12**, 43–51 (1976)
5. MARCUS, R. S., KUGEL, P. and BENENFELD, A. R. Catalog information and text as indicators of relevance, *Journal of the American Society for Information Science* **30**, 25–30 (1978)
6. LEMON, N. *Attitudes and Their Measurement*, Batsford, London (1973)
7. ZIPPERER, W. C. User interface models for multidisciplinary information dissemination systems, ERIC Microfiche ED122846 (1976)
8. SARACEVIC, T. Relevance: a review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science* **26**, 321–343 (1975)

9. LESK, M. E. and SALTON, G. Relevance assessments and retrieval system evaluation, *Information Storage and Retrieval* **4**, 343–359 (1968)
10. REES, A. M. and SCHULTZ, D. G. *A Field Experimental Approach to Relevance Assessments in Relation to Document Searching*, 2 Vols, Centre for Documentation and Communication Research, Case Western Reserve University (1967)
11. MILLER, G. A. *The Psychology of Communication*, Basic Books, New York (1967)
12. GILBERT, H. and SPARCK JONES, K. *Statistical Bases of Relevance Assessment for the 'Ideal' Information Retrieval Test Collection*, British Library Research and Development Report 5481, Computer Laboratory, University of Cambridge (1979)
13. COOPER, M. D. Input–output relationships in online bibliographic searching, *Journal of the American Society for Information Science* **28**, 153–156 (1977)
14. SPARCK JONES, K. and VAN RIJSBERGEN, C. J. Information retrieval test collections, *Journal of Documentation* **32**, 59–75 (1976)
15. TAGUE, J. and FARRADANE, J. Estimation and reliability of retrieval effectiveness measures, *Information Processing and Management* **14**, 1–16 (1978)
16. HARTIGAN, J. A. *Clustering Algorithms*, Wiley, New York (1975)
17. SALTON, G. *Dynamic Library and Information Processing*, Prentice-Hall, Englewood Cliffs, N.J. (1975)
18. WINER, B. J. *Statistical Principles in Experimental Design*, 2nd edn, McGraw-Hill, New York (1971)
19. FISHER, R. A. and YATES, F. *Statistical Tables for Biological, Agricultural, and Medical Research*, Oliver and Boyd, Edinburgh (1953)
20. KERLINGER, F. N. *Foundations of Behavioral Research*, Holt, Rinehart and Winston, New York (1964)
21. SPARCK JONES, K. Performance averaging for recall and precision, *Journal of Informatics* **2**, 95–105 (1978)
22. MENDENHALL, W. W. *et al. Statistics: a Tool for the Social Sciences*, Duxbury Press, North Scituate, Mass. (1974)
23. WINKLER, R. L. and HAYES, W. L. *Statistics: Probability, Inference, and Decision-Making*, 2nd edn, Holt, Rinehart and Winston, New York (1975)
24. KENDALL, M. G. and STUART, A. *The Advanced Theory of Statistics*, Vols 1–3, Hafner, New York (1961–1966)
25. NOETHER, G. E. *Elements of Nonparametric Statistics*, Wiley, New York (1967)
26. GILCHRIST, W. *Statistical Forecasting*, Wiley, New York (1976)
27. BOX, G. E. P. and JENKINS, G. W. *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco (1976)
28. MOORE, P. W. *Introduction to the Use of Computer Packages for Statistical Analysis*, Prentice-hall, Englewood Cliffs, N.J. (1978)