Section B : Statistical methods of determining relevance assessment requirements

This section constitutes the main body of the Report. It is devoted primarily to the detailed presentation, in technical terms, of the statistical methods considered for determining relevance assessment requirements. The presentation is intended to be comprehensive and self-contained. But since it is recognised that some at least of these interested in the 'ideal' collection may not be statistically trained, an attempt has been made to provide separate, brief, Noddy-style accounts of the arguments as prefaces to the full presentation of each main method*. One of the methods, the 'Pool' method, was initially put forward in the design study report. This is therefore examined first. The other methods investigated, and in particular the main alternative 'Squares' method, are then described, and the Pool and Squares methods are compared. These methods are designed to provide assessments for future experiments comparing two indexing or searching strategies, and the section concludes with a discussion of approaches to assessment for multi-strategy comparisons. The structure of the section is therefore as follows.

B1 Preface: the research proposed and its conduct

The application was for a small grant to support a trained statistician for three months to carry out the following pieces of work:

---

* There is thus a certain amount of repetition, but it was thought that this was acceptable in the interests of self-contained non-specialist and specialist presentations.

a) to approach the relevance assessment problem from first principles i) to check the initial approach of the design study report (Sparck Jones and Bates (1977)) and either confirm it or present an alternative; and ii) to work out the consequent detailed figures for the numbers of assessments required in different circumstances. This was envisaged as analytic work based on the assumptions originally made, and in particular on two strong assumptions, namely that at building time the searches could be guaranteed to draw out all the relevant documents for a request, and all the documents which would be retrieved in any sensible search.

b) to develop the argument to cover the cases where these assumptions are weakened, as they certainly should be, i.e. to allow for only most (some specified percentage) of the relevant or possible output documents being retrieved. This development could perhaps be pushed through analytically, or could involve some fairly straightforward computer simulation using standard statistical packages. Again detailed figures should be derived.

In general, where appropriate, tests with available collections should be conducted to check the statistical arguments.

The work has been carried out by H. Gilbert, a graduate student from the University Statistical Laboratory, with advice from Dr. P. Altham of that Laboratory, and from Dr. S.E. Robertson and Dr. C.J. van Rijsbergen. In general the conduct of the project has been as proposed: the main difference has been that a range of methods has been considered, even though the study of the initial method confirmed the arguments on which it is based.

The project work can be split into two parts. Firstly, a detailed examination was made of the method of the design study report, here referred to as the Pool method. This examination took the form of tightening the assumptions made and improving the accuracy of the figures produced. Secondly, alternative methods were considered in the hope of finding one which reduces the number of assessments required. Of these alternatives two are presented in detail, one of which was abandoned for reasons stated in chapter B3 and the other, referred to as the Squares method, is presented as a viable alternative in chapter B4. The search for a suitable method of obtaining assessments involved a fairly extensive literature survey, so the discussion of approaches to assessment in this report is hopefully a comprehensive one.

Any lengthy tables have been relegated to Appendix 1, while the computing which the work involved is discussed in Appendix 2.

B1.1 General requirement

The main criterion used when calculating the sample size required to be assessed is that it should be large enough to make a comparison between any two retrieval strategies meaningful.

Meaningful has a specialised interpretation here which is centred on the statistical concepts of significance level and power.

Suppose that we wish to test the hypothesis that there is no difference in recall between strategies A and B for a set of requests against the alternative hypothesis that strategy A is better. Denote the first hypothesis, the null hypothesis, by H_o. Then there are two possible mistakes which could be made

(i) H_o could be rejected when it is, in fact, true

(ii) H_o could be accepted when it is false.

Clearly it is desirable in any statistical test to place an upper bound on the probabilities of these events occurring. The significance level of a test refers to the upper bound on the probability of event (i) occurring. Equivalent to placing an upper bound on the probability of event (ii) occurring is placing a lower bound on the probability of rejecting H_o when it is false. This probability is known as the power of the test.

The calculations in the remainder of section B are based on a significance level of 0.05(or 0.01) and a power of 0.95.

In general this means that the probability of making an incorrect decision has been reduced to an acceptable level.

For example, it will be shown later (see Table 1) that if we have 300 requests and an average of 25 relevant or retrieved documents per request, then if we wish the probability of error(i) (and the probability of error(ii)) occurring to be less than 0.05, there must have been 15 documents of known relevance status assessed and the number of successes (defined as A better than B) must be greater than 167. To achieve the 15 known documents 60% of the post must have been assessed.

B2 The Pool method

B.2.1 Non-statistical summary

The method discussed below is referred to for convenience as the Pool method. The argument for it is basically the same as that used in the design study report, with some alterations and improvements.

One of the major changes from the original argument is that the strong assumptions concerning the output of possible future searches and the percentage of relevant documents contained in the pool are no longer necessary. The cost of discarding these assumptions is that instead of the argument resulting in what percentage of the pool should be assessed per request, it results in saying how many documents should be assessed given that the pool size is N. That is, for each different size of pool a separate calculation must be performed to reveal the number of documents to be assessed.

A rough, non-statistical outline of the argument is as follows:

suppose that there are k requests in the request set and that we wish to compare strategies A and B for this set. Then these k requests are thought of as k trials and in each trial the recall (or alternatively precision) value of strategies A and B are compared. To compare two strategies only the part of each output which has been assessed is considered. Note that the recall value is the probability of a document being retrieved given that it is relevant and the precision value is the probability of a document being relevant given that it is retrieved.

Next it is noted on how many trials the recall (precision) value of A is better than the corresponding value of B. This number is then tested to see if it is likely to have arisen just through the variation due to sampling or whether it represents a real difference between the strategies. If the latter is the case then A is said to be better than B.

Clearly, for the comparison between A and B (typically in future experiments using the 'ideal' collection) to be valid on this basis, the assessed sample must be adequate. This assessment sample would be drawn from

the pooled output of searches done when the collection was built, and in the design study report it was assumed that the pool would contain the output of any future strategies and all relevant documents for a request. The pool could thus be quite large, but for increasingly large request sets it was found that progressively smaller samples would need assessment. The argument to establish the percentage sample required for a given request set size was essentially as follows. It works backwards from the way, described above, the assessments are to be used.

We assume that what we are trying to establish is that there is a significant difference between two probabilities (or two proportions), based on sample estimates of them. That is we wish to show that there is a significant difference between recall, or alternatively precision, expressed as a probability, for two strategies A and B, given the results of applying A and B for a set of requests; the difference itself is to be analysed in terms of probabilities, respectively that strategy A is better than B (Prob(A>B)) and B better than A (Prob(B>A)). The sets of search results are regarded as samples because, as we cannot have exhaustive assessments, we compare A and B for each request with respect to a collection subset consisting of documents of known relevance status. (Notice that this will ordinarily be a subset of the sets of documents actually retrieved by A and B.) We therefore need to know how large this evaluation sample must be for a valid comparison between A and B for any request and hence over all requests, and further to know what size of assessment sample of the pool is required to ensure that the outputs of any strategies A and B respectively will contain large enough valuation samples. Different sizes of evaluation and assessment sample will be required for different sizes of request set and of germane document set, as described below. It must moreover be emphasised that while the formal argument for sample size is the same for recall and for precision, the evaluation sample for recall must consist only of relevant documents, while for precision the documents may be either relevant or non-relevant. The expression "documents of known relevance status" will be used to cover both cases where a specific distinction is not required, but the underlying difference between the two should be borne in mind.

For the purposes of discussion we regard an occurrence of A>B as a success. To characterise the probability distribution of successes we use the normal approximation to the binomial distribution. As a significance test to be applied

to our comparison between A and B we choose the sign test because it makes few a priori assumptions about the data. For the two strategies we order each request in terms of effectiveness, i.e. the effectiveness of a request under A $\gtrless$ its effectiveness under B. Effectiveness here is either precision or recall, which are regarded as probabilities. The null hypothesis, H_o, is that there is no difference between A and B for the set of requests, i.e. Prob(A>B) = Prob(B>A) = ½. Since the test is based on the binomial distribution we can use the normal approximation in an entirely standard way to find the critical region for the test, that is, that value of the standardised normal variable which needs to be exceeded for H_o to be rejected at 5% significance level. If k is the number of requests, and x the number of successes, then under H_o : Prob(A>B) = Prob(B>A) = ½ and we get

$$\frac{2x - k}{\sqrt{k}} \sim N(0,1)$$

Using normal tables we then find

$$\frac{2x - k}{\sqrt{k}} > 2$$

gives 5% significance. This means, for example, that if k = 100 (requests), we must have at least 60 successes, i.e. 60 requests where A>B for either recall or precision, whichever we are using.

Proceeding as just described would be all that was necessary if there were no uncertainties about the probabilities being compared, that is no uncertainty about recall or precision for each request, because strategy output could be related to global assessment information. Unfortunately it cannot, as this information is lacking, and we are in fact trying to decide whether A>B or B>A on the basis of two samples, one for A and one for B. Because requests differ, the difference between A and B will vary across the request set. So even if there is a real difference between A and B, it will be obscured, and even more so if the evaluation samples for the requests are unreliable due to size, or to bias in the assessment samples from which they are derived. Clearly if the samples for the requests are infinite (very large), the difference between A and B over the set can be confidently established; but such sampling is exactly what is not feasible.

We therefore first assume that the probabilities we are trying to estimate, i.e. recall or precision, are constant over requests, that is that

the number or relevant documents or of retrieved documents are the same for all requests. This is clearly artificial, but may be interpreted from a practical point of view as referring to averages. We can then calculate the minimum evaluation sample size for each (and hence every) request necessary for the sign test to show a significant difference. This in turn requires an assumed real difference between the two strategies. The bigger the real difference, the smaller the sample size needed to reflect it; however prudence suggests assuming a small real difference, which may be taken, conventionally, as 5%, i.e. $p_A - p_B = 5\%$. For the calculation a standard sampling theorem for differences, again allowing the use of the normal approximation to the binomial distribution, can be utilised. For a given real difference of 5% and some given (evaluation) sample size n, the theorem gives us $P(x_A > x_B)$, i.e. the probability of recall (or precision) for A being greater than that for B. Conversely, for given $P(x_A > x_B)$ we can derive the sample size n required to achieve P. Then as constancy across requests is being assumed, we can obtain, for P, the expected number of requests with A>B; or conversely, if we take as our expected number that number required by the sign test, as described above, we can work backwards to determine the sample size needed to achieve this expected number of requests with A>B. Thus for the example above with 300 requests, and an expected number of requests with A>B = 167, we would need an evaluation sample of 15 documents of known relevance status.

However if we design our data for this expected proportion of successes, this is not sufficient because of the uncertainty introduced by the fact that we are sampling. That is, while our evaluation sample may in principle be adequate to tell us whether we have the required number of successes, in practice we cannot rely on accurate enough data to obtain the required evaluation sample in our output more than half the time. If we provide ourselves with a larger evaluation sample, supplying in principle more information than we need, we have a better chance of obtaining enough actual information in practice. In other words, referring back to the discussion of the significance test, in designing for the expected number, we may in practice find that the number of A's>B's will fall below the critical value represented by this number 50% of the time. We would in fact like a higher chance of significance in our results, i.e. a higher chance of rejecting the null hypothesis if it is indeed false (so $p_A - p_B = 5\%$ is true). This can only be done by increasing the probability of $P(x_A > x_B)$, which means increasing

the evaluation sample size. Specifically we want to ensure a 95% chance when $p_A - p_B = 5\%$ that the number of A's>B's exceeds the critical value. We thus ask for what value $P(x_A > x_B)$ will it be the case that there is a 95% chance of significance rather than only a 50% chance. We use the normal approximation to the binomial, as before, to obtain this value, and hence the required size of evaluation sample.

Once we have the sample size we can use it in a straightforward way to calculate the percentage of the pool to be assessed, and so can draw a random assessment sample from each request pool as appropriate. This is done by relating the number of specific documents with known relevance status of the evaluation sample to the presumed or known total number of documents associated with a request. For the computation of recall the latter is the total of relevant documents, for precision the total of retrieved documents for any strategy. Table 0 gives an illustrative selection of the somewhat crudely calculated figures of the design study report, to show how the argument works out in practice, throughout interpreting the constant figures of the statistical argument as averages. For example we see that for 5% significance and 500 requests, given 9 known relevance status documents and a total of 50 relevant or retrieved documents per request, 18% of the pool needs assessment; for 300 requests and 15 known, 30% needs assessment. Thus if the pool obtained for requests in building the 'ideal' collection contained 1000 documents, this would imply assessment of a random sample of 180 documents in the first case and 300 in the second.

In practice slightly different consequences ensue according to whether we are interested in recall or precision, because we are concerned with different document sets, the relevant and the retrieved, respectively, and these are normally different. If they were identical, as above, the same percentage of the pool would have to be assessed. However as the former is usually smaller this implies, for a given known relevant sample, a higher percentage assessment. Recall thus imposes more stringent requirements for assessment than precision, and should so be used in considering collection building effort.

As will be shown in the technical discussion, paragraph 2.2.2, the essential argument for the Pool method can be developed without the use of the two assumptions about the nature, but not size of the pool, but as a tradeoff a specific pool size must be given.

For reference the complete set of assumptions underlying the whole argument as originally presented in the design study report may be summarised as follows.

1 for future experiments comparing strategies A and B

1 we evaluate using recall and precision;
2 recall and precision are probabilities estimated by proportions based on samples;
3 we use the sign test for validating performance differences;
4 a percentage difference, say of 5%, between A and B, in recall or precision, is indicated by $Prob_A - Prob_B = 5\%$;
5 a normal sampling distribution for difference of proportions;
6 a normal approximation to the binomial distribution for the power of the sign test.

2 for assessment data

1 all relevant documents are contained in the pool;
2 the output of A, and of B, is contained in the pool;
3 a sample from the pool is a random sample;
4 a pool random sample is also a strategy output random sample.

3 for request data in evaluation and assessment

1 the requests are independent;
2 the probability of finding strategy A better than strategy B is constant across requests.

The detailed analysis which follows confirms that the original argument for the Pool method was essentially correct, and that the detailed figures given in the design study report were relatively accurate. More carefully worked figures were derived, and are given in Table 1. With respect to the assumptions given above, those about the evaluation methodology were perforce retained. For the assessment data, the strong assumptions 2.1 and 2.2 could, as mentioned, be jettisoned. Overall, the really important assumptions are those about the requests, 3.1 and 3.2, and these cannot be avoided.

B2.2 Statistical presentation

B2.2.1 Scrutiny of design study argument

The experimental design assumed was that each of the two strategies was applied to a set of requests. The sign test was then used. That is, suppose there are k requests: then let n_i be the number of assessed documents relevant to the i^{th} request, a_i and b_i be the number of documents relevant to the i^{th} request retrieved by strategies A ans B respectively ($i = 1,2,..,k$), and let n_A, n_B be the number of documents retrieved by strategies A and B respectively.

The recall and precision values for strategy A can then be estimated by the proportions $\frac{a_i}{n_i}$ and $\frac{a_i}{n_A}$ respectively. So A is considered to be better than B for the i^{th} request (using the recall criterion) if $\frac{a_i}{n_i} > \frac{b_i}{n_l}$ (that is, $a_i > b_i$).

Define the random variable X_i ($i = 1,..,k$) such that

$$X_i = \begin{cases} 1 \text{ if } a_i > b_i \\ 0 \text{ otherwise.} \end{cases}$$

Then the sign test refers $\sum_{i=1}^{k} X_i$ to $Bi(k,\frac{1}{2})$. (1)
The reason for this is that, if the statement H_o is assumed to be true (H_o = that there is no difference between the strategies), then $P(a_i > b_i)$ (= probability that $a_i > b_i$) = $P(b_i > a_i) = \frac{1}{2}$ for $i = 1,..,k$.

From (1) the probability of any specific value of ΣX_i occurring, given that H_o is true, can be obtained. If this probability is below the significance level of the test (and is therefore unlikely to have occurred due to sampling variation), then H_o is rejected in favour of A being better than B if ΣX_i is large, and in favour of B being better than A if ΣX_i is small.

Having decided on using the sign test the next stage was to find out the minimum number of requests for which strategy A would have to be better than strategy B in order to reject H_o in favour of A being better than B (at the 5% [or 1%] significance level).

The way that this was done in the design study was to use the normal

approximation to the binomial distribution, that is:

$$\text{as } k \to \infty \text{ , } \frac{x - kp}{\sqrt{kp(1 - p)}} \sim N(0,1) \qquad (2)$$

where x is the number of successes (that is, the number of requests on which A was better than B), and p is the probability of success (that is, $P(a_i > b_i)$ for recall), which is assumed to be constant over the request set.

An improvement which was made here was to introduce the correction for continuity. This correction is necessary since the normal distribution is for a continuous variable while the binomial distribution involves a discrete variable. Regarding the observed frequency x as occupying an interval, the lower limit of which is half a unit below the observed frequency while the upper limit is half a unit above the observed frequency, the correction consists of reducing by 0.5 the difference between the observed value of x and the expected value of x (kp).

So (2) is replaced by

$$\frac{(x \pm 0.5) - \frac{1}{2}k}{\sqrt{k/4}} \sim N(0,1) \quad .$$

(Note that $p = \frac{1}{2}$ since we are assuming that H_o is true.) Clearly A would not be deemed better than B if $x < \frac{1}{2}k$, so it follows that x can be assumed to be larger than $\frac{1}{2}k$ and so we have

$$z = \frac{(x - 0.5) - \frac{1}{2}k}{\sqrt{k/4}} \sim N(0,1) \quad .$$

Using the tables of the normal distribution (Cambridge Statistical Tables) it can be seen that H_o is rejected in favour of A being better than B at the 5% significance level if

$$\frac{x - 0.5 - \frac{1}{2}k}{\sqrt{k/4}} > 1.96 \quad .$$

That is

$$x > \frac{1.96\sqrt{k} + k + 1}{2} \quad .$$

Similarly H_o would be rejected at the 1% level if

$$x > \frac{2.58\sqrt{k} + k + 1}{2} \quad .$$

This calculation provides us with column three of Table 1 (see Appendix 1).

Now that an upper bound on the probability of mistakenly rejecting H_o has been established we turn to placing a lower bound on the probability of correctly rejecting H_o, that is, on the power of the test. In this particular case 0.95 is the lower bound which was chosen. That is, we now have to find the value of p (=p_o, say) such that

$$P(x > \frac{1.96\sqrt{k + k + 1}}{2} / p) \geq 0.95, \forall p \geq p_o \quad .$$

Define

$$x_c = [\frac{1.96\sqrt{k + k + 1}}{2}]$$

(the integral part); then, since x can only take integer values we require p_o such that

$$P(x > x_c / p) \geq 0.95, \forall p \geq p_o \quad .$$

That is,

$$P(z > \frac{x_c - 0.5 - kp}{\sqrt{kp(1 - kp)}} / p) \geq 0.95, \forall p \geq p_o$$

where $z \sim N(0,1)$.

This value can be found by experimenting with different values of p and using a set of tables of the standardised normal distribution. Hence the fourth column of Table 1 is obtained.

Column five lists the values of the sample size required to identify a difference between the two strategies with 95% power and 5% (1%) significance. To obtain this column we need to use the following sampling theorem which can be found in Hoel (1971), p. 135. Since the pool is being sampled, the observed values of the recall (precision) proportion for strategies A and B are random variables. So let $\hat{p}_A$ and $\hat{p}_B$ represent the observed values based on n_A and n_B trials respectively, for a particular request, from two binomial populations with probability p_A and p_B respectively. Then p_A and p_B denote the true recall (precision) values for strategies A and B. Applying the central limit theorem we obtain:

Theorem

When the number of trials n_A and n_B are sufficiently large (say, >25) or $\hat{p}_A$ and $\hat{p}_B$ are normally distributed with mean $\mu_{\hat{p}_A - \hat{p}_B} = p_A - p_B$ (note that $p_A - p_B$ denoted the true difference in recall [precision] between strategies A and B) and variance

$$\hat{\sigma}^2_{\hat{p}_A - \hat{p}_B} = \frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B} \quad ,$$

where $q_A = 1 - p_A$. In the recall case $n_A = n_B$ = the number of documents

which are relevant to the particular request, while in the case of precision n_A and n_B are the number of documents retrieved by strategies A and B respectively.

In the argument of the design study report it was then tacitly assumed that $n_A = n_B = n$, say. This assumption clearly holds for recall since the number of documents relevant to any particular request is independent of which strategy is searching for them (remembering that we are dealing with documents which have already been assessed). However in the case of precision the assumption is saying that the number of assessed documents retrieved by each strategy is the same, which is clearly not a realistic assumption.

The assumption can be avoided by defining $n = \min(n_A, n_B)$. We now define

$$z = \frac{(\frac{a_i}{n_A} - \frac{b_i}{n_B}) - (p_A - p_B)}{\sqrt{(\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B})}}$$

where a_i and b_i were defined on page B10. Then $z \sim N(0,1)$ (by the theorem). Then we must choose n such that $P(a_i > b_i) \geq p_o$ (see column four of Table 1) so that we have 95% power. Unless we assume a lower limit on the real difference in recall (precision), that is $p_A - p_B$, we are unable to do this, so assume that $p_A - p_B \geq 0.05$.

We enlarge here upon the argument for precision, since the argument for recall is the same (with $n_A = n_B$).

Now

$$P(\frac{a_i}{n_A} - \frac{b_i}{n_B} > 0) = P(z > \frac{-0.05}{\sqrt{(\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B})}})$$

But we do not know the values of p_A and p_B, so we must use the inequality obtained by simple calculus, that $p_A q_A \leq \frac{1}{4}$, $p_B q_B \leq \frac{1}{4}$. This inequality is not too crude, since for $\frac{1}{4} < p < \frac{3}{4}$, pq only varies over the range (0.187, 0.250).

Therefore, if

$$P(z > -0.5\sqrt{2n}) \geq p_o \qquad (3)$$

then

$$P\left(z \geq \frac{-0.05}{\sqrt{\left(\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}\right)}}\right) \geq p_o \quad .$$

Using the tables of the normal distribution we can find the smallest value of n satisfying (3). If the recall criterion is being used, n is equivalent to the number of assessed relevant documents, and if we are considering precision then n is the number of assessed documents retrieved by a strategy.

If exhaustive relevance judgement was possible then the above argument would be sufficient, with n denoting the number of documents of known relevance status per request. That is, given the number of requests available, the above argument obtains the minimum number of relevant documents in the case of recall, or retrieved documents in the case of precision, which must be assessed so that the result holds at the 5% significance level. This calculation was carried out in detail and Table 2 and Figure 1 were constructed.

Table 2(a) indicates how many requests are required so that a result can be obtained at the 5% (1%) significance level with 95% power, given the minimum number of documents of known relevance status (minimised over the request set).

The reverse requirement, how many documents are required for a given number of requests, is shown in Table 2(b). For example, if there are 200 requests, then unless each request has at least 21 documents of known status a result cannot be obtained at the 5% significance level.

Note that these figures only concern those experiments where the sign test is used to analyse the results. An alternative to the sign test is presented in chapter B4.

In the theorem above it was stated that $\text{var}(\hat{p}_A - \hat{p}_B) = \frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}$. McNemar (1947) points out that this only holds when the samples from which the proportions $\hat{p}_A$ and $\hat{p}_B$ are drawn from are independent. In this

case however, since the two strategies search the same set of documents, there is a definite correlation between $\hat{p}_A$ and $\hat{p}_B$. This correlation results in a decrease in the variance, which would mean that the values of n obtained in column five of Table 1 are slightly on the large size.

B2.2.2 Improved sampling rationale: the modified Pool method

In the argument used in the design study report, after n is calculated, the percentage of the pool required to be assessed in order to have, say, n documents of known relevance status assessed is calculated. The way this was done was to say that, for recall, if one has K relevant documents altogether in the set and it is required to have n assessed then, if it is assumed that all the relevant documents are in the pool, one should estimate $100 \times \frac{n}{K}$ % of the pool. For the precision criterion the parallel argument assumes that all the output of future searches is contained in the pool. However, when the specified proportion is assessed then very often n required documents would not be obtained in the sample, since n is only the number which one would expect to obtain.

An alternative method which gives more chance of obtaining the required number of documents of known relevance status is the following. Putting it in terms of recall, suppose that we have observed N documents in the pool, K of which are relevant, and suppose that at least n of these K must be assessed. Then, if a simple random sample of size S is taken for assessment, the probability that at least n relevant documents are contained in this sample is

$$\frac{\sum_{y=n}^{\min(S,K)} \binom{K}{n}\binom{N-K}{S-n}}{\binom{N}{n}} \qquad \text{(hypergeometric distribution)}$$

where $\binom{N}{n} = \frac{N!}{(N-n)!n!}$.

Therefore S can be chosen such that the probability of having assessed n relevant documents is at least 0.95. Table 3 lists the values of S for various values of N, K and n. In order to construct the table, a program was written in standard FORTRAN (see Appendix 2). (Note that the level of confidence is a user-controlled parameter of the program. We illustrate the results of setting it to 95% but it could of course be set lower.)

Note that this argument does not require any assumptions about the output of future strategies or whether or not all the relevant documents are in the pool; also, one can say with greater confidence that the required number of relevant documents have been assessed. For reasons given in Appendix 2, there is small error in each of the figures, but it is still safe to use them provided the lower bound of 0.95 is taken as approximate (0.93 - 0.97).

As an example, consider the case when there are 300 requests and the significance level is taken to be 5%. Then Table 1 states that 15 relevant documents are required to be assessed, and that if there are 25 relevant documents altogether then 60% of the pool should be assessed.

However, if there are 1000 documents in the pool and 600 of them are assessed, then Table 3 states that we can only be "95% confident" that 11 relevant documents have been assessed, and that at least 729 documents would have to be assessed to be equally confident of assessing 15 relevant documents.

B2.2.3 Accuracy of estimators

As was mentioned earlier, $\hat{p}_A$ and $\hat{p}_B$ are observed estimates of the true proportions p_A and p_B. If these estimates are required to have a certain accuracy then this places another lower bound on the number of relevant documents to be assessed for recall or the number of documents retrieved for precision.

So suppose that n_A is to be chosen to ensure that

$$P(|\hat{p}_A - p_A| > d) \leq \alpha \qquad (4)$$

where d and α are small and

$$\hat{p}_A \sim N(p_A, (1 - \frac{n_A}{N})\frac{N}{N-1}\frac{p_A q_A}{n_A}) \quad .2.16 \text{ Barnett, p.39.}$$

Assume, for the moment, that $\frac{n_A}{N}$ (the sampling fraction) is small enough that $1 - \frac{n_A}{N}$ can be approximated by 1, and that N is large enough that $\frac{N}{N-1}$ is close to 1. Then

$$z = \frac{\hat{p}_A - p_A}{\sqrt{\frac{p_A q_A}{n_A}}} \sim N(0,1) \quad .$$

So n_A must be chosen such that

$$P(|z| > \frac{d}{\sqrt{\frac{p_A q_A}{n_A}}}) \leq \alpha \quad .$$

That is,

$$\frac{d}{\sqrt{\frac{p_A q_A}{n_A}}} \geq z_\alpha$$

where z_α is the double-tailed α-point of N(0,1). That is, $P(|z| > z_\alpha) = \alpha$, and it can be found from the tables. So

$$n_A \geq \frac{p_A q_A z_\alpha^2}{d^2} \quad .$$

The problem now is that p_A and q_A are unknown, but we know that $p_A q_A$ has a maximum value of ¼, so taking

$$n_A \geq \frac{z_\alpha^2}{4d^2}$$

will certainly satisfy the accuracy requirements of (4).

If the sampling fraction is large enough that it has to be retained then

$$\text{var}(p) = (1 - \frac{n_A}{N})\frac{Np_A q_A}{(N-1)n_A} \quad .$$

So to satisfy $P(|\hat{p}_A - p_A| > d) \leq \alpha$ we need

$$n_A \geq \frac{p_A q_A}{v}(1 + \frac{1}{N}[\frac{p_A q_A}{v} - 1])^{-1} \qquad (5)$$

where

$$v = (\frac{d}{z_\alpha})^2 \quad .$$

(5) is not directly applicable since p_A is not known precisely. Again this can be overcome by replacing $p_A q_A$ by ¼. These calculations were performed in detail and tabulated in Table 4.

Note that whenever we have used the inequality $p_A q_A \leq$ ¼ alternative forms of action could be

1) perform a pilot study to yield a preliminary estimate of p_A, or
2) earlier experimental work using strategy A may give an indication of the true value of p_A in this case.

B2.2.4 Weakening the pool assumptions

In paragraph B2.2.1 of the foregoing the Pool method was presented using the strong assumptions of the design study report to the effect that the pool contains all relevant and retrieved documents. In paragraph B2.2.2 it was shown that these assumptions could be abandoned at the cost of having some specific pool size, i.e. knowing the pool size and calculating the assessment sample accordingly. The effect of weakening the assumptions without abandoning them altogether was therefore investigated to see what this would imply for assessment. Table 5 gives the percentage of the pool required if only 90% of the relevant documents (or of the search output) is deemed to be in the pool. Weakening the assumptions presents no difficulties for the general structure of the argument, and does not have any very drastic effects on the numbers of documents requiring assessment, especially for larger numbers of relevant or retrieved documents. For example, for recall for 300 requests and 25 relevant documents, assuming only 90% coverage in the pool implies a 66.7% assessment sample rather than a 60% one. In general weakening the assumptions to the indicated extent means that the percentage of the pool to be assessed increases by up to 10 percent, which in practice would probably not be too costly.

B2.2.5 Conclusion on the Pool method

The argument presented in the design study report is statistically correct given that the assumptions hold; but the way in which the percentage of the pool required for assessment is calculated can be improved on, as indicated. Further, the original assumption set can be reduced: specifically the assumptions about the content of the pool are not required. The important assumptions which are still required are therefore:

3.1 the requests are independent;

3.2 $P(a_i > b_i)$ is constant for $i = 1,..,k$.

Assumption 3.2 that $P(A > B)$ is constant across requests is clearly unlikely to be true, and if it were not then our calculations for column three of Table 1 would be invalid, although for the moment this assumption is sufficient to enable us to obtain an idea of the magnitude of the sample size required to be assessed.

Also, in the case when there are only 5 or 10 documents of known

relevance status the sampling theorem cannot be used unless it is assumed that $\hat{p}_A$ and $\hat{p}_B$ are normally distributed. How confidently this assumption can be made is a question which anyone using the sign test with a small number of documents of known status should consider carefully.

As assumption 3.2 is somewhat unrealistic, alternative methods not requiring it or other equally strong assumptions were sought. The investigation of alternatives was also stimulated by the desire to simply reduce the number of assessments required, in order to lower the cost of building the 'ideal' collection. The lines of work undertaken are described in the following chapter.

B3 Other methods: dead ends

This section is concerned mainly with methods which were investigated which turned out turned out to be unsatisfactory. One of the main difficulties was the fact that most tests applicable are conditional but in this situation no data is available, since we require to know the sample size to be assessed before the search strategies are applied.

The first approach was a rather lengthy literature survey to see if this problem(or an analogous one) had been considered elsewhere. This proved to be virtually fruitless. In particular the type of paper which was looked at was that dealing with 2×2 contingency tables, since these were thought to be the most informative way of tabulating data. A typical example of such a table for the present retrieval context would be the following.

	retrieved by A	not retrieved by A	
retrieved by B	a_i	c_i	$a_i + c_i$
not retrieved by B	b_i	d_i	$b_i + d_i$
	$a_i + b_i$	$c_i + d_i$	n_i

(1)

where a_i is the number of relevant documents (for the i^{th} request) retrieved by strategies A and B, b_i is the number of relevant documents retrieved by strategy B but not by strategy A, etc. An analogous table

could be constructed for non-relevant documents.

B3.1 Logistic model

One way of approaching the problem using this technique is to construct such a table for each request and to be left with a series of 2×2 contingency tables. Cox (1970) analyses a series of 2×2 tables resulting from two independent samples. However, as was pointed out earlier, here the samples are not independent (since both strategies search the same document set), and so Cox's method could not be applied directly. To overcome this we made use of the logistic model which is outlined below.

Suppose there are k requests, and let n_i denote the number of documents relevant to the i^{th} request.

Define

$$x_{ij} = \begin{cases} 1 \text{ if the } j^{th} \text{ document relevant to the } i^{th} \text{ request} \\ \text{is retrieved by strategy A} \\ 0 \text{ otherwise} \end{cases}$$

$$i = 1,..,k,\ j = 1,..,n_i,$$

and

$$y_{ij} = \begin{cases} 1 \text{ if the } j^{th} \text{ document relevant to the } i^{th} \text{ request} \\ \text{is retrieved by strategy B} \\ 0 \text{ otherwise} \end{cases}$$

$$i = 1,..,k,\ j = 1,..,n_i \quad .$$

Then assume the following model

$$\log \frac{P(x_{ij} = 1)}{P(x_{ij} = 0)} = \delta + \lambda_{ij},\ j = 1,..,n_i,\ i = 1,..,k$$

$$\log \frac{P(y_{ij} = 1)}{P(y_{ij} = 0)} = -\delta + \lambda_{ij},\ j = 1,..,n_i,\ i = 1,..,k,$$

which assumes that the difference between A and B is constant (on the logistic scale) across requests. Note that the parameter δ can be thought of as the 'strategy' effect and λ_{ij} can be thought of as the 'document' effect.

So, it follows that

$$P(x_{ij},y_{ij}) = \frac{e^{(\delta + \lambda_{ij})x_{ij}}}{(1 + e^{\delta + \lambda_{ij}})} \frac{e^{(-\delta + \lambda_{ij})y_{ij}}}{(1 + e^{-\delta + \lambda_{ij}})}$$

(assuming x_{ij}, y_{ij} are independent).

So

$$P(\underline{x},\underline{y}\ \delta,\lambda_{ij}) = \prod_{i=1}^{k}\prod_{j=1}^{n_i} P(x_{ij},y_{ij} \,/\, \delta,\lambda_{ij})$$

$$= \frac{e^{\delta(x.. - y..)} e^{\Sigma\Sigma\lambda_{ij}(x_{ij} + y_{ij})}}{\prod_{i=1}^{k}\prod_{j=1}^{n_i} (1 + e^{\delta + \lambda_{ij}})(1 + e^{-\delta + \lambda_{ij}})}$$

where $x.. = \sum_{i,j} x_{ij}$, $y.. = \sum_{i,j} y_{ij}$.
Therefore the distribution of $(x.. - y..)/(x_{ij} + y_{ij},\ i = 1,..,k,\ j = 1,..,n_i)$ depends only on δ. So we should consider the distribution of $x..$ $(x_{ij} + y_{ij},\ i = 1,..,k,\ j = 1,..,n_i)$. So we seek the distribution of the sum of independent variables.

If $x_{ij} + y_{ij} = 0$ the $x_{ij} = 0$, and if $x_{ij} + y_{ij} = 2$ then $x_{ij} = 1$, so it is sufficient to consider the distribution of

$$\sum_{\substack{i,j: \\ x_{ij}+y_{ij}=1}} x_{ij} \,/\, (x_{ij} + y_{ij},\ i = 1,..k,\ j = 1,..,n_i) \quad .$$

This is equivalent to considering the distribution of

$$\sum_{1}^{k} b_i \,/\, (b_i + c_i,\ i = 1,..,k) \qquad \text{(see (1))}.$$

It is easily seen that

$$P(x_{ij} = 1 \,/\, x_{ij} + y_{ij} = 1) = \frac{e^{\delta}}{e^{\delta} + e^{-\delta}} \quad ,$$

and so, on the null hypothesis that there is no difference in performance between the strategies ($\delta = 0$) we obtain that

$$b_i \,/\, (b_i + c_i) \sim Bi(b_i + c_i, \tfrac{1}{2}) \quad .$$

Given $(b_i + c_i,\ i = 1,..,k)$, $b_1,..b_k$ are independent; thus

$$\sum_{1}^{k} b_i \,/\, (b_i + c_i,\ i = 1,..,k) \sim Bi\left(\sum_{1}^{k}(b_i + c_i), \tfrac{1}{2}\right) \quad .$$

So, with this model, for a uniformly most powerful unbiassed size α test, we should reject $\delta = 0$ in favour of $\delta > 0$ (that is, A better than B) if and only if

$$\sum_{1}^{k} b_i > C\left(\sum_{1}^{k}(b_i + c_i), \alpha\right)$$

where C is found from $Bi\left(\sum_{1}^{k}(b_i + c_i), \tfrac{1}{2}\right)$ tables (probably a normal approximation will do).

Note that if

$$A_i = \begin{cases} 1 \text{ for } b_i > c_i \\ 0 \text{ otherwise} \end{cases}$$

$i = 1,..,k$

and we do a sign test, treating $\sum_{1}^{k} A_i$ as $Bi(k, \tfrac{1}{2})$ on the null hypothesis, we

presumably get a less powerful test than the one above, since the parameters of the binomial distribution are respectively k and $\sum_{1}^{k}(b_i + c_i)$ and by definition $k < \sum_{1}^{k}(b_i + c_i)$. So the sign test will be conservative.

However, in practice the value of $\sum_{1}^{k}(b_i + c_i)$ is not known before the experiment. There are two main ways of trying to overcome this.

Firstly, choosing a small sample and thus obtaining a confidence interval for $\sum_{1}^{k}(b_i + c_i)$. Alternatively, the expected value of $\sum_{1}^{k}(b_i + c_i)$ can be used. Now with x_{ij}, y_{ij} and b_i defined as before, $i = 1,..,k$, $j = 1,..,n_i$ it is easily seen that

$$b_i = \sum_{j=1}^{n_i} x_{ij}(1 - y_{ij})$$

so

$$E(b_i) = \Sigma \frac{e^{\delta + \lambda_{ij}}}{j1 + e^{\delta + \lambda_{ij}}} - \Sigma \frac{e^{2\lambda_{ij}}}{j(1 + e^{\delta + \lambda_{ij}})(1 + e^{-\delta + \lambda_{ij}})} .$$

Similarly,

$$c_i = \sum_{j=1}^{n_i} (1 - x_{ij})y_{ij}$$

and so

$$E(c_i) = \Sigma \frac{e^{-\delta + \lambda_{ij}}}{1 + e^{-\delta + \lambda_{ij}}} - \Sigma \frac{e^{2\lambda_{ij}}}{(1 + e^{\delta + \lambda_{ij}})(1 + e^{-\delta + \lambda_{ij}})} .$$

Hence

$$E(b_i + c_i) = \Sigma \frac{e^{\delta + \lambda_{ij}}}{j1 + e^{\delta + \lambda_{ij}}} + \frac{e^{-\delta + \lambda_{ij}}}{1 + e^{-\delta + \lambda_{ij}}} \quad \frac{2e^{2\lambda_{ij}}}{(1 + e^{\delta + \lambda_{ij}})(1 + e^{-\delta + \lambda_{ij}})}$$

and so $E(\Sigma(b_i + c_i))$ follows easily, and clearly depends on λ_{ij}, $i = 1,..,k$, $j = 1,..,n_i$. These parameters are also unknown, and so this is as far as the analysis of this method reached. One remaining possibility was to assign a prior distribution to the λ_{ij}'s (for example $\frac{e^{\lambda}}{1 + e^{\lambda}}$ uniform on (0,1)), and then perform a Bayes preposterior analysis. However, the integrations involved would have to be done numerically and would be multidimensional. Also, because of the number of parameters involved, it is hard to see what is being said, implicitly, about the data when a particular prior distribution is assigned to the parameters; so it is doubtful whether or not the time spent performing the integrations would have been worthwhile. This method was therefore abandoned so that alternatives could be examined.

B3.2 Wilcoxon's signed ranks test

The next method to be attempted was replacing the sign test of the design study report by Wilcoxon's matched-pairs signed-ranks test. So the situation is that the two strategies, A and B, are compared, as before, over the request set.

Now let d_i denote the difference in recall (precision) between the strategies for the i^{th} request. It now has to be decided whether or not it is meaningful to rank these differences: if so, let r_i denote the rank of d_i. Suppose there are k requests in total; then define

$$T_w = \sum_1^k r_i s_i$$

where $s_i = +1$ if strategy A is better than strategy B for the i^{th} request and -1 otherwise. Assuming hypothesis H_o to be true, it is seen that $s_i = +1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. Thus $E(s_i) = 0$ and so $E(T_w) = 0$, and it follows that $var(T_w) = E(T_w^2)$.

Now

$$E(T_w^2) = E(\sum_1^k r_i s_i)^2 = E(\sum r_i^2 s_i^2 + 2 \sum_{i \neq j} r_i s_i r_j s_j) \quad .$$

It can easily be shown that

$$E(s_i^2) = 1 \text{ and } E(s_i s_j) = 0 \qquad (i \neq j),\ i,j = 1,..,k.$$

Therefore

$$var(T_w) = E(T_w^2) = \sum_1^k r_i^2 = \frac{1}{6}k(k+1)(2k+1) \quad .$$

Now denote P(A > B) (again assumed constant over the requests) by p. Then

$$E(s_i) = 2p - 1,\ i = 1,..,k$$

and so

$$E(T_w) = \sum_1^k r_i(2p-1) = \tfrac{1}{2}k(k+1)(2p-1) \quad .$$

$$E(T_w^2) = \sum_1^k r_i^2 E(s_i^2) + 2 \sum_{i \neq j} r_i r_j E(s_i s_j) \quad .$$

$E(s_i^2) = 1$ as before. Assuming independence between requests it follows that

$$E(s_i s_j) = E(s_i)E(s_j) = (2p-1)^2 \quad .$$

Thus

$$E(T_w^2) = \tfrac{1}{6}k(k+1)(2k+1) + 2(2p-1)^2 \sum_{i \neq j} r_i r_j \quad ;$$

$$2 \sum_{i \neq j} r_i r_j = (\Sigma r_i)^2 - \Sigma r_i^2 = \tfrac{1}{4}k^2(k+1)^2 - \tfrac{1}{6}k(k+1)(2k+1) \quad .$$

Therefore

$$\text{var}(T_w) = E(T_w^2) = \tfrac{2}{3}k(k+1)(2k+1)p(1-p) \quad .$$

For k large enough (greater than 25) we can assume that T_w is normally distributed, and since (given p) its mean and variance are known, the exact form of the distribution is known.

Table 6 is analogous to Table 1. Column four lists the smallest value c, say, such that

$$P(|T_w| > H_o) \leq 0.05$$

and was obtained by using the fact that under H_o

$$T_w \sim N(0, \tfrac{1}{6}k(k+1)(2k+1))$$

and by using tables of the normal distribution. This results in

$$c = (\tfrac{1}{6}k(k+1)(2k+1))^{\frac{1}{2}}(1.96) \qquad \text{(5\% significance)}$$

Column five lists the smallest $p > 0.5$ such that

$$P(T_w > c / p) \geq 0.95$$

and this is calculated in the same manner as in paragraph B2.2.1. Similarly column six is arrived at by the same means as before.

It should be noted that the number of documents of known relevance status required to be assessed is higher than when the sign test was used, and so the sign test is better. This is surprising since Wilcoxon's matched-pairs test would appear to be making fuller use of the data as it assigns more weight to a request which shows a large difference between the two conditions than to a request which shows a small difference (that is, it pays attention to the magnitude of the difference as well as the direction).

The fact that the sign test is better in this situation can be proved algebraically as follows.

The power of the sign test and of Wilcoxon's test are

$$P\left(z > \frac{c^* - kp}{\sqrt{kp(1-p)}}\right)$$

and

$$P(z > \frac{c - \frac{1}{2}(2p - 1)k(k + 1)}{\sqrt{\frac{2}{3}k(k + 1)(2k + 1)p(1 - p)}})$$

respectively, where $z \sim N(0,1)$, $c = \sqrt{(\frac{1}{6}k(k + 1)(2k + 1))}(1.96)$ and $c^* = \frac{(1.96)\sqrt{k} + k + 1}{2}$.

So the sign test is more powerful than the Wilcoxon matched-pairs signed-ranks test if

$$\frac{c - \frac{1}{2}(2p - 1)k(k + 1)}{\sqrt{(\frac{2}{3}k(k + 1)(2k + 1)p(1 - p)}} > \frac{c^* - kp}{\sqrt{kp(1 - p)}} \quad .$$

After some simple algebraic manipulation this reduces to

$$(2k\sqrt{2k + 1} - k\sqrt{6(k + 1)})p > (k + 1)\sqrt{2k + 1} - k\sqrt{\frac{3(k + 1)}{2}} \quad .$$

The table below lists the minimum value of P(A > B) (that is, p) for the sign test to be more powerful, given the number of requests, k.

k	p
300	0.513
400	0.509
500	0.507
600	0.506
700	0.505
800	0.505
900	0.504
1000	0.504

As, in general, the value of p required for the test to have 95% power is greater than 0.55, it follows that the sign test will be more powerful.

This result is probably due to the fact that the underlying distribution is binomial. Apart from the fact that more documents of known status are required to be assessed, other disadvantages of Wilcoxon's matched-pairs signed ranks test are the cost involved in ranking the differences and also that the ranking may not be meaningful. So this method is not recommended.

B3.3 Likelihood ratio test

Next the likelihood ratio test was attempted. For this it was assumed that

$$y_i = \hat{p}_{A_i} - \hat{p}_{B_i} \sim N(p_A - p_B, \frac{p_A q_A + p_B q_B}{n_i})$$

where $\hat{p}_{A_i} - \hat{p}_{B_i}$ are the observed recall values for the i^{th} request, and we test

$$H_o : p_A - p_B = 0 \quad vs \quad H_1 : p_A \neq p_B.$$

Thus we have a random sample $(y_1,..y_k)$ where k is the number of requests.

The likelihood function

$$P(\underline{y} ; \theta) = \prod_{i=1}^{k} \frac{1}{\sqrt{\frac{2\pi m}{n_i}}} \exp(\frac{-n_i}{2m}(y_i - (p_A - p_B))$$

where $\theta = (p_A,p_B)$, n_i is the number of assessed documents relevant to the i^{th} request, and $m = p_A q_A + p_B q_B$ (assuming independence of the requests and that p_A, p_B remain constant over the request set).

We consider the ratio

$$\frac{P(\underline{y} / \hat{\theta},H_1)}{P(\underline{y} / \theta_o,H_o)}$$

and reject H_o if it is too large. First $P(\underline{y} / \theta,H_1)$ has to be maximised with respect to p_A and p_B. Equivalently, maximise $\log P(\underline{y} / \theta,H_1)$. This results in having to solve the following equation

$$2kp_A^3 - (3k + N)p_A^2 + p_A(2kp_B(1 - p_B) - 2\Sigma n_i y_i - 2p_B N - k) + p_B(1 - p_B) + \Sigma_1 n_i y_i + p_B N - kp_B(1 - p_B)$$

where $N = \Sigma n_i$, and a similar equation with p_A and p_B interchanged. This method was not pursued.

B4 The Squares method

B4.1 Non-statistical summary

This section is devoted to an alternative method which could be developed in a satisfactory way. This method, which we have called the Squares method, is therefore the main competitor to the Pool method, and in chapter B5, following the technical presentation of the Squares method, the two are compared.

The Squares method is based on a retrieval contingency table giving, for two (indexing, searching) strategies A and B, the number of documents retrieved by both A and B, the number retrieved by A but not by B, by B but not A, and by neither A nor B: i.e. the table cells represent A ∧ B, A ∧ -B, -A ∧ B, and -A ∧ -B. Such a table may in principle refer to any documents retrieved, or more specifically to relevant documents only, or to non-relevant documents. In the present context the first of these is irrelevant: our interest is in the ability of strategies to obtain relevant documents, or avoid non-relevant ones. Further, as the object is to compare A and B, the cells of interest are those where the two strategies differ, i.e. those where A retrieves but B does not, or vice versa. It is moreover evident, as the tables are confined to either relevant or non-relevant documents, that the method is suited to the use of recall as a performance measure, but not precision, for which tables would have to be combined, and fallout is therefore substituted for precision as a performance measure. The discussion which follows is carried out in terms of recall, the treatment of fallout being easily seen by analogy.

Essentially the Pool method aims at determining, within the framework given by the use of the sign test, specified significance level, etc., whether the evaluation sample of documents of known relevance status justifies the assertion that strategies A and B differ in performance. The Squares method approaches the problem from a rather different angle by looking to see what differences in the specific contingency table cells are required to support the assertion; i.e. it is concerned to determine whether, for recall superiority in strategy A say, the number of relevant documents retrieved by A alone, in relation to the number of relevant documents retrieved by either A or B alone, is larger than would be expected if A and B performed the same. In other words, we evaluate A ∧ -B/((A ∧ -B) ∨ (-A ∧ B)). Alternatively if the number of documents retrieved by A is sufficiently smaller than would be expected, the null hypothesis that A and B do not differ is rejected in favour of B's superiority to A.

The method requires certain assumptions. In order to forecast the required sample size, i.e. total number of known relevant documents, a value has to be assigned (a) to the probability that a relevant document is retrieved by only one strategy and (b), more specifically (assuming that we are looking for A better than B), to the probability that a relevant

document is retrieved by A, given that it is retrieved by just one strategy. We may deal with (b) by placing a lower bound on the probability, analogous to the assumption $p_A - p_B \geq 5\%$ for the Pool method. But for (a) a specific value is desirable, so if this is not forthcoming from previous experience, for example, the only course is to set a lower bound here too and take this as the value the probability has.

The appropriate sample size is obtained as follows. If n_i is the total of known relevant documents for request i, a_i the relevant documents retrieved by A and B, b_i those retrieved by A alone, c_i those retrieved by B alone, and d_i those retrieved by neither, we have a contingency table as follows:

		B: retr	B: not retr	
A	retr	a_i	b_i	
	not retr	c_i	d_i	
				n_i

Summing over k requests we define $n = \sum_i^k n_i$, $a = \sum_i^k a_i$, b..., c... and d... similarly, to obtain Table I:

relevant documents

		B: retr	B: not retr	
A	retr	a	b	
	not retr	c	d	
				n

I

We first assume that the distribution of (a,b,c,d), given the value of n, is multinomial with probabilities (p_a, p_b, p_c, p_d), defined as $Mn(n; p_a, p_b, p_c, p_d)$; i.e. we assume that the probability of a relevant document not being retrieved by either strategy is p_d, of being retrieved by B alone is p_c, by A alone is p_b, and by both is p_a. Thus Table I has a corresponding probability table:

relevant documents		B: retr	B: not retr
A	retr	p_a	p_b
A	not retr	p_c	p_d

II

Given then some known relevant documents, the probability of obtaining some specific (observed) set of values for a,b,c,d in Table I, denoted by P(a,b,c,d/n), is

$$P(a,b,c,d/n) = \begin{cases} \frac{n!}{a!b!c!d!} p_a^a p_b^b p_c^c p_d^d & \text{if } a + b + c + d = n \\ 0 \text{ otherwise} & . \end{cases}$$

The recall value for strategy A is determined by $p_a + p_b$ (in practice estimated by $\frac{a + b}{n}$ and for strategy B it is $p_a + p_c$, so to compare strategies A and B for recall it is sufficient to compare p_b and p_c.

Now if the value of b + c is known the distribution of b is $Bi(b + c, \frac{p_b}{p_b + p_c})$, that is has a binomial distribution. There are b + c relevant documents which could land in the (1,2)th element of Table I, and for any such document the probability that it does land there is $\frac{p_b}{p_b + p_c}$, or Δ say. So, given the value of b + c, and using the normal approximation to the binomial distribution (as for the evaluation of the critical region for the Pool method), we can find $K(\alpha, b + c)$ such that the hypothesis $p_b = p_c$ is rejected in favour of $p_b > p_c$ just when $b > K(\alpha, b + c)$. K is a constant depending on the significance level α of the test used (one assuming a multinomial distribution for the table), and on the value of b + c. If $p_b = p_c$ then $\Delta = \frac{1}{2}$, so b has the distribution $Bi(b + c, \frac{1}{2})$. Thus if $\alpha = 0.05$, say, K must be such that $P(b > K / p_b = p_c) < \alpha$.

As for the Pool method, we require stable power of the test as well as a specified significance level; specifically, we want the power of the test to be at least 0.95. This imposes a lower bound on the value of Δ, which can be found by trial and error, like the lower bound imposed on P(A > B) for the Pool method. In practice this means that when the true value of Δ exceeds this lower bound we can be confident that we are correctly rejecting the hypothesis $p_b = p_c$, i.e. the power is greater than 0.95.

Unfortunately, the value of b + c is not known (in advance, that is, of actual experiments with any pair of strategies A, B). However it is known that b + c has the distribution $Bi(n,\Pi_R)$, where Π_R represents $p_b + p_c$ (the probability a document is retrieved by one strategy only), so if some value of Π_R is assumed, the exact form of its distribution will be known and we can then find r and s, say, such that $P(r \leq b + c \leq s) = 0.95$. We can also find the expected value of s, say e. Then given any specific number of relevant documents, n in Table I, and assuming some specific value of Π_R, substituting b + c by r, e and s respectively gives the value of $K(\alpha,b + c)$, and the lower bound on the true value of Δ. If the nature of the strategies being compared is such that they may be expected to retrieve different relevant documents, the assumed value of Π_R should be quite large, while if they are expected to be the same, the assumed value should be low.

The Pool method specifies requirements for the number of documents to be assessed for each request, to provide enough relevance information for that request. The Squares method is concerned only with sets of documents of known relevance status for a set of requests. This is an advantage for the latter,as will be discussed more fully in chapter B5, However it should be noted here that relevance assessments are necessarily assessments for individual requests and the only way of obtaining the total set n is by summing the assessments for the individual requests. Thus the Squares method necessarily relies on a pool and sample basis for assessment like that used for the Pool method. In other words the argument assumes that n is obtained as the sum of n_i's, where each n_i is a random sample of a comprehensive pool. However, as will be emphasised in chapter B5, as there is no requirement that the individual n_i's satisfy specific requirements, and should not vary, the size of sample and of request set can be chosen pragmatically, and the latter in particular can be adjusted in relation to the observed rather than expected properties of the n_i's and accumulating n.

To illustrate the kind of figure obtained from the argument (see Table 7), for 5% significance and 95% power in the test, and a conservative $\Pi_R = 0.25$, i.e. a value of Π_R assuming a high overlap in output between strategies A and B and hence a small performance difference between them: if n = 3000, b + c has a lower confidence bound of 704 and b must exceed 378 if A is to be accepted as better than B, where the minimum value of Δ

required for 95% power must be ≥ 0.568. On the other hand, if $\Pi_R = 0.50$, assuming less overlap, n = 2000, b + c has lower bound 938, b must exceed 499, and Δ must ≥ 0.559. If we are concerned with recall, the more awkward case, we could perhaps expect to get a total of 2000 relevant documents from 80 requests averaging 25 relevant documents each, and 3000 from 120 requests. We need a large n as this is associated with a low value of Δ: a low Δ represents a realistic expectation about the number of documents in cell b in the contingency table in relation to those in b + c, i.e. about the number of documents to be retrieved by strategy A alone, as opposed to by either A or B alone. Thus if n = 50, the expected value of b + c = 12.5, b is at least 11 and $\Delta = 0.935$; and the chance in practice of finding such a relatively large b is very poor.

The assumptions underlying the Squares method can be summarised, for comparison with those used for the Pool method, as follows

1 <u>for future experiments comparing strategies A and B</u>

1 we evaluate using recall and fallout;
2 recall and fallout are probabilities estimated by proportions based on samples;
3 the distribution of the retrieved document sets a,b,c,d in the contingency table comparing A and B, conditional on the total n, is $Mn(n; p_a,p_b,p_c,p_d)$;
4 the distribution of set b (retrieved by A but not B), conditional on the set b + c retrieved by A alone or B alone, is $Bi(b + c,\frac{p_b}{p_b + p_c})$ (alias $Bi(b + c,\Delta)$);
5 the distribution of b + c is $Bi(n,p_b + p_c)$ (alias $Bi(n,\Pi_R)$);
6 a normal approximation to the binomial distribution for the power of the significance test.

2 <u>for assessment data</u>

the same four assumptions as the Pool method.

3 <u>for request data in evaluation and assessment</u>

1 the requests are independent.

The two methods thus share assumptions 2.1 - 2.4, and 3.1. But as the technical argument of paragraph B2.2.2 above showed, assumptions 2.1 and 2.2 are not necessary and can be abandoned in favour of simply knowing the pool size (though the argument was presented under the Pool method it is equally applicable to the Squares method). Assumptions 1.1 and 1.2 have the same general character for the two methods, with the specific difference that fallout replaces precision. Assumption 1.6 for the Squares method is

also like 1.6 for the Pool method. On the other hand an important feature of the Squares method is that the strong assumption 3.2 required for the Pool method, namely that the probability of finding strategy A better than strategy B is constant across requests, is not required. The distinctive asssumptions for the Squares method, 1.3, 1.4 and 1.5, replacing 1.3 - 5 for the Pool method, follow from the structure imposed on the data, and lead to a rather different argument. However within the framework of this argument these assumptions are somewhat analogous to those used for the Pool method.

B4.2 Statistical presentation

We wish to compare the two strategies A and B. Let n_i denote the total number of assessed documents relevant to the i^{th} request, a_i denote the number of these retrieved by both strategies, b_i the number retrieved by strategy A but not strategy B, c_i denote the number retrieved by strategy B but not by strategy A, and d_i denote the number retrieved by neither strategy. This information can be summarised in the following table:

relevant documents		B: retr	B: not retr	
A: retr		a_i	b_i	
A: not retr		c_i	d_i	
				n_i

The Squares method presented below can only be used if retrieval data can be set out in this form.

We now define, for k requests

$$n = \sum_{i=1}^{k} n_i,\quad a = \sum_{i=1}^{k} a_i,\quad b = \sum_{i=1}^{k} b_i,\quad c = \sum_{i=1}^{k} c_i \text{ and } d = \sum_{i=1}^{k} d_i ,$$

and obtain Table I:

relevant documents		B: retr	B: not retr	
A: retr		a	b	
A: not retr		c	d	
				n

I

Clearly, to compare the two strategies for the request set, a and d are not as important as b and c. This method bases its decision on the size of b relative to b + c: if b is relatively large (small) then the hypothesis of equality can be rejected in favour of strategy A being better than strategy B (vice versa).

Table I has an analogous table of probabilities, namely

relevant documents		B: retr	B: not retr
A	retr	p_a	p_b
	not retr	p_c	p_d

II

and tables analogous to Tables I and II exist for non-relevant documents, namely

non-relevant documents		B: retr	B: not retr	
A	retr	a'	b'	
	not retr	c'	d'	
				n'

non-relevant documents		B: retr	B: not retr
A	retr	$p_{a'}$	$p_{b'}$
	not retr	$p_{c'}$	$p_{d'}$

.

Let N denote the total number of documents assessed, and suppose that the distribution of (a,b,c,d/n) in $Mn(n; p_a,p_b,p_c,p_d)$ and that the distribution of (a',b',c',d'/N - n) is $Mn(N - n; p_{a'},p_{b'},p_{c'},p_{d'})$. That is,

$$P(a,b,c,d/n) = \begin{cases} \frac{n!}{a!b!c!d!} p_a^a p_b^b p_c^c p_d^d & \text{if } a + b + c + d = n \\ 0 \text{ otherwise} \end{cases} .$$

Similarly P(a',b',c',d'/N - n).

Now the recall value of strategy A is $p_a + p_b$ and for strategy B it is $p_a + p_c$. So strategy A is better for recall if $p_b > p_c$. Similarly strategy A is better for fallout if $p_{b'} < p_{c'}$. (Note that strategy A is better

for precision if

$$\frac{p_a + p_b}{p_a + p_b + p_{a'} + p_{b'}} > \frac{p_a + p_c}{p_a + p_c + p_{a'} + p_{c'}} \quad ;$$

however, as using precision for evaluation requires a combination of relevant and non-relevant document tables, and generally leads to complexity, it is much more convenient to use fallout to complement recall, as fallout only requires the one non-relevant document table.

The conditional test of $p_b = p_c$ against $p_b > p_c$ has the rule, reject $p_b = p_c$ if and only if

$$b > K(b + c,\alpha) \quad . \tag{1}$$

That is, reject $p_b = p_c$ if b is larger than a constant K, which depends on $b + c$ and α (the significance level). Also the distribution of $b / (b + c)$ is $Bi(b + c , \frac{p_b}{p_b + p_c})$ (providing $b + c \neq 0$), so the conditional power of the test is

$$P(b > K(b + c,\alpha) \; / \; b + c, \frac{p_b}{p_b + p_c}) \tag{2}$$

and the unconditional power of (1) is

$$E_{b+c}(P[b > K(b + c,\alpha)] \; / \; b + c, \frac{p_b}{p_b + p_c}) \quad .$$

It is known that $b + c \sim Bi(n, p_b + p_c)$ and so

$$P(b + c = k) = \binom{n}{k}(p_b + p_c)^k(1 - p_b - p_c)^{n-k} \quad .$$

Therefore the power is

$$\sum_{b+c=1}^{n} P(b > K(b + c,\alpha) \; / \; b + c,\Delta)P(b + c)$$

where $\Delta = \frac{p_b}{p_b + p_c}$.

So in order to find the power of the test $P(b + c = k)$ must be evaluated for $k = 1,..,n$ and $K(b + c,\alpha)$ must be found, that is the smallest K such that

$$P(b > K \; / \; b + c, \Delta = \tfrac{1}{2}) \leq 0.05$$

where $b \sim Bi(b + c, \tfrac{1}{2})$.

Since n will usually be very large the normal approximation to the binomial distribution can be used to calculate $P(b + c)$; also for $b + c > 10$

the normal approximation can be used to find K.

Finally $P(b > K(b + c,\alpha) / \Delta)$ has to be calculated for $b + c = 1,..,n$. Again the normal approximation can be used for $b + c > 10$.

Since n is likely to be very large these calculations should clearly be performed on a computer. However precisely because n is very large and $P(b + c)$, $K(b + c,\alpha)$ and $P(b > K(b + c,\alpha) / \Delta)$ must all be calculated for $b + c = 1,..,n$, it would take an extremely large amount of computer time to find the power for any particular values of n and Δ.

As part of the object of this report is to construct fairly comprehensive tables, which in this case means varying the values of n and Δ, it was decided that it would not be feasible to calculate the exact power on a computer.

A rather crude method of overcoming this problem was to calculate a 95% confidence interval for b + c and then to replace b + c in (2) by its upper and lower bounds in the confidence interval, and by its expected value. This results in three different values for the conditional power of the test,which should give us a reasonable indication of its unconditional power.

Now $b + c \sim Bi(n,\Pi_R)$ where $\Pi_R = p_b + p_c$, so $E(b + c) = n\Pi_R$. Also we wish to find r and s such that

$$P(r \leq b + c \leq s)\left(= \sum_{j=r}^{s} \binom{n}{j} \Pi_R^{\,j} (1 - \Pi_R)^{s-j} \right) = 0.95 \quad .$$

That is, to find the largest r and smallest s such that

$$P(b + c \leq r) \leq 0.025 \text{ and } P(b + c \geq s) \leq 0.025.$$

Again we use the normal approximation, which leads us to choose the largest r and smallest s such that

$$\frac{r - n\Pi_R}{\sqrt{n\Pi_R(1 - \Pi_R)}} \leq -1.96 \text{ and } \frac{s - n\Pi_R}{\sqrt{n\Pi_R(1 - \Pi_R)}} \geq 1.96.$$

So a value of Π_R (the probability in relation to recall that a relevant document is only retrieved by one strategy) must be assumed. Once a value for b + c is found (viz r, s or E(b + c)), the fact that

$$b \sim Bi(b + c, \tfrac{1}{2})$$ (the distribution is conditional on b + c)

if the strategies have no difference in performance, is used to evaluate K such that H_o is rejected if b > K.

It is now necessary to ensure that the test has 95% power, and this forces a lower bound on Δ. This can be found by a trial and error method similar to the one used to find the lower bound of p in the Pool method.

Table 7 lists the value of K for various values of n and Π_R, and also gives the minimum value of $\frac{p_b}{p_b + p_c}$ for which one can be confident (in the sense of achieving 95% power) of correctly rejecting H_o. For each value of n and Π_R the values of b + c tabulated are on the lower bound of the 95% confidence interval, its expected value, and the upper bound of the confidence interval respectively.

Predictably enough the minimum value of $\frac{p_b}{p_b + p_c}$ increases as n and/or Π_R increases. Also as n increases there is little difference between the three values of the lower bound for fixed values of Π_R.

Before using Table 7 an estimate must be made of Π_R. If the two strategies are very different then Π_R may be thought to be very high (≥ 0.75, say), while if the two strategies are very similar than Π_R could be about 0.25 (or less) (Π_R is expected to be small if b + c is expected to be small relative to n, and is expected to be large if b + c is expected to be relatively large, since, if the experiment had been performed then Π_R would be estimated by $\frac{b + c}{n}$).

Suppose that two strategies A and B are being compared for recall and that it is expected that approximately half the relevant documents will be retrieved by only one strategy (that is $\Pi_R \sim 0.50$) (information concerning Π_R may be available from earlier studies). If one has a total of 2000 relevant documents then reject H_o (at the 5% significance level) if b > 563 (assuming the 'worst' case for b + c). One then has 95% power of correctly

detecting that A is better than B if $\frac{p_b}{p_b + p_c} \geq 0.559$.

Note that since b + c is expected to lie between 938 and 1062, one has to take the worst case both for the value of K and for the lower bound on Δ.

Conversely, suppose that one wishes to be confident (in the sense of 95% power and 5% significance) of deciding that A is better than B, whenever $\Delta \geq 0.55$; then approximately 3000 documents (relevant to any of the requests) are required to be assessed (assuming $\Pi_R = 0.50$). Note that there may well be overlap in these documents,but providing the requests are assumed independent this does not matter. Further, this overlap in documents does not affect the number of assessments made. The hypergeometric program of Appendix 2 assumes that all the documents are distinct, but it is equally applicable when they are not, since the assessments are all distinct.

Finally note that A is better than B for fallout if $p_{b'} < p_{c'}$. So we can use the same argument as for recall, only this time n is the total number of non-relevant documents.

Summary

This test is a conditional test and so, before being able to apply it a value of Π_R and a lower bound on Δ must be assumed. Independence between requests is assumed but this time P(A > B) is not assumed to be constant across requests. The actual mechanics of choosing the sample size could prove a problem: this is discussed in more detail in the next chapter.

B5 Comparison between the Pool and Squares methods

This chapter compares the Pool and Squares methods, primarily from the point of view of their interpretation and implications.

To make actual numerical comparisons is difficult since the number of documents to be assessed if the Squares method is used depends on Π_R. Comparisons when Π_R is assumed are made later in the chapter.

One of the main differences between the methods is the amount of information which must be obtained. For the Pool method all that is needed is how many relevant documents the individual strategies retrieved. However for the Squares method a further classification is required, since as well as knowing the above, one must also know the number of documents retrieved by strategy A but not by strategy B, and vice versa. This is obviously a disadvantage since presumably the increase in information supplied can only be achieved by some increase in costs.

Another difference is that the Pool method depends much more heavily on the number of requests than the Squares method. In the former the number of documents needing to be assessed descreases as the number of requests increases. However in the latter the dependence is not so explicit since here the power of the test increases as b + c increases. Clearly b + c depends on the number of requests, but it also depends on how general the requests are, since the more general the requests are presumably the larger b + c will be (since there will be more relevant or retrieved documents). Therefore it is conceivable that the same value of b + c could occur for a set of 200 requests and a set of 250 requests. In this situation there is an increase in power in the Pool method but the power remains unchanged in the Squares method. So, for the Squares method, the value of b + c is more important than the number of requests.

Whichever method is used there is a problem connected with the sampling aspect.

In the Pool method we are interested in the number of documents of known relevance status each request has. If we have k requests the method supplies us with the number of documents (n_i) which must be assessed, for the i^{th} request, in order to make the comparison 'meaningful' (as defined earlier), for i = 1,..,k. The problem is that the number of documents varies considerably from request to request (see chapter A2). However we can only make one assessment of the pool and this must result in a sufficient number of documents of known status for each request (some of which may not even have n_i documents, especially relevant ones, in the pool).

It would be unduly pessimistic to choose the size of the sample to be assessed on the basis of the requests with the fewest relevant documents.

A more practical method would be to take the average number of documents of known status per request, as if all requests have that number, and sample accordingly.

For example, to take the recall case, suppose there are 3 requests with 25, 50 and 75 relevant documents respectively, and the Pool method requires 20 relevant documents per request. Then if we want the probability of achieving this number to be at least 0.95 in each request then over 80% of the pool must be sampled because of the small number of documents relevant to the first request. However, this would presumably result in assessing over 60 documents relevant to the third request and so the actual power of the test is greater than 0.95, and a lot more resources than necessary have been used. So, as a compromise suppose that each request has 50 relevant documents and then assess a significantly smaller percentage of the pool (< 60%).

A question which is raised in this connection is whether requests with a small number of relevant documents should be discarded, but this introduces some bias into the request set (see Section C for further discussion on this point).

The situation is much simpler for the Squares method since we are no longer interested in n_i but in n ($= \Sigma n_i$), that is, in the total number of relevant documents. This therefore removes all problems of variation between requests and makes the sampling problem much simpler. The only drawback here, as mentioned in the last chapter, is the fact that though the n documents are unlikely to be distinct they have to be treated as distinct as they have to be specifically assessed in relation to their various requests. In other words, n refers to the total of relevant document postings, and there can be no economising on assessment if the total of different documents retrieved is smaller.

However, even if there is this slight defect, overall there is no doubt that the Squares method does not create as many sampling problems as the Pool method and, in this way, has a big advantage.

Both methods also require some assumptions to be made about the size of the difference in performance one wishes to be confident about detecting.

In the Pool method this amounts to assuming that $p_A - p_B \geq 0.05$, and this assumption was made when the sampling theorem was used (see chapter B2 ($p_A = p_a + p_b$, $p_B = p_a + p_c$, in Table II.

In the Squares method this assumption took the form of choosing a lower bound for $\frac{p_b}{p_b + p_c}$ (i.e. Δ). Thus if, for example, it is assumed that $\Delta \geq 0.55$, then in order to achieve 95% power and 5% significance one needs (to the nearest 100) a total of 5200 documents to be assessed, and specifically 5200 relevant documents to be identified for recall evaluation.

The assumption made by the Pool method has the advantage of being easily interpreted since $p_A - p_B$ is the difference in recall while $\frac{p_b}{p_b + p_c}$ is the probability of a document of known relevance status being retrieved by strategy A given that it is only retrieved by one strategy.

Suppose now that strategies A and B are being compared for recall, and that there are 300 requests, 1000 documents in the pool, and an average of 25 relevant documents per request; and that we require 5% significance and 95% power. Then there are 7500 relevant documents altogether.

Then, using the Pool method and applying Table 1 in conjunction with Table 3 we obtain that 729 documents should be assessed. Before using the Squares method, a value of $p_b + p_c$ (i.e. Π_R) must be assumed (given that we are interested in $\Delta \geq 0.55$, say). If $\Pi_R = 0.25$ then we require 5200 relevant documents to be assessed. The hypergeometric distribution tells us that, in order to have, with 95% probability, assessed 5200 relevant documents, more than 729 documents must be assessed. However if $\Pi_R = 0.50$ then we only need to have assessed 3000 relevant documents, and so the number of assessments required is certainly less than 729. Clearly things become even better if $\Pi_R = 0.75$.

So whether or not the number of assessments required is less for the Squares method than for the Pool method depends on the value of Π_R. If Π_R is quite low then the Pool method will probably be better, while if Π_R is large the Squares method becomes more efficient. That is, if the total number of documents retrieved or rejected by <u>both</u> strategies, i.e. the overlap

in their outputs, is likely to be relatively large then the Pool method will require less assessments than the Squares method.

So the Squares method has the advantages of a simpler sampling problem and of requiring less assessments than the Pool method when the strategies are not too similar, as in general could not be expected, while the Pool method does not require as much information from the data as the Squares method, has a more meaningful constraint, and is the better method when the strategies are very similar. On the other hand the Squares method does not have to assume the constancy of P(A > B) across requests, which the Pool method does.

B6 Multi-strategy comparisons

Up till now we have only considered considered comparisons between pairs of strategies. What happens if one wishes to compare t strategies, $A_1, A_2, .., A_t$ taken together say, which is a reasonable requirement for retrieval experiments?

There are three obvious ways of approaching this problem. The first two, Cochran's Q-test and Friedman's test, are standard non-parametric tests, while the third is based on David (1963). The first two are quite independent of the previous methods; David's method is somewhat similar to the use of the sign test.

Cochran's Q-test

Cochran (1950) has shown that if there is no difference in, say, recall under each strategy, then if k (the number of requests) is not too small

$$Q = \frac{k(k-1)\sum_{j=1}^{k}(G_j - \bar{G})^2}{k\sum_{1}^{N}L_i - \sum_{1}^{N}L_i^2}$$

is distributed approximately as chi-square with k - 1 degrees of freedom where G_j denotes the number of relevant documents retrieved by strategy A_j and L_i denotes the total number of strategies which retrieve the i^{th} relevant document.

So, the null hypothesis that the strategies have the same level of

performance is rejected at the 5% significance level if Q is in the upper or lower 2.5% tail of the chi-square distribution with k - 1 degrees of freedom.

Note however that the power of Cochran's test is not known exactly, and so no forecasts can be made about the size of the sample taken for assessment. Another disadvantage is the fact that rejecting H_o does not result in an ordering of the strategies. Also, Cochran's test would be rather expensive to put into practice since one would have to note, for each relevant document, howe many strategies retrieved it.

Friedman's two-way analysis of variance

First of all the data must be expressed in a two-way table having k requests and t strategies. If there is no difference between the performance of the strategies then, providing the number of requests and/or strategies is not too small ($k > 9$, $t > 4$) it can be shown (Friedman (1937)) that χ_r^2 is distributed approximately as chi-square with k - 1 degrees of freedom when

$$\chi_r^2 = \frac{12}{kt(t+1)} \sum_{j=1}^{t} (R_j)^2 - 3k(t+1)$$

where R_j denotes the sum of the ranks in the j^{th} column.

Again the power of the test is unknown. Siegel (1956) points out that the test compares favourably with the F-test and so should be used in preference to Cochran's Q-test.

Method derived from David (1963): pair subset comparison

Cochran's and Friedman's methods are unsatisfactory in that they do not supply enough quantitative information about the power of the test, and hence what their future use would imply for the asessment data cannot be determined. The method for pair subset comparison described below can be applied to future strategy comparisons, if multiple comparisons are wanted, and does not impose any requirements on the assessment data. It is simply applied to judgements of the relative merits of pairs of strategies.

This method was first considered as a possible alternative to the Pool method (in the case where t = 2). It considers the case when there are t objects to be compared in pairs by k different judges. So this can be applied to

our problem with the strategies replacing "objects" and requests replacing "judges". Each judgement is assumed to consist of saying merely which object is best and ties are not permitted.

Suppose t = 2, and denote the number of preferences scored by A_i (i = 1,2) by a_i. Since under the hypothesis H_o of equality strategy A_i has probability ½ of being preferred in each of its k comparisons with the other strategy, the score a_i is a binomial $Bi(n,\frac{1}{2})$ variate. So the method is equivalent to the use of the sign test in the Pool method.

For the case when t > 2 this method would appear to be a viable alternative to Cochran's and Friedman's tests. For more details see David (1963).

If it is decided to impose an ordering on the strategies by the application of David's method, one must beware of circular triads occurring. That is, $A_1 > A_2$, $A_2 > A_3$, $A_3 > A_1$. These could result from the fact that there may be no valid ordering of the three strategies since their performance may depend on more than one characteristic. Also if there is not a significant difference between the strategies then the comparisons (if no ties are allowed) cannot reasonably be expected to be consistent. If the sign test (Pool method) is being used then one can decrease the chances of this by combining the multinomial model and the sign test as follows.

Consider the table for the r^{th} request:

Relevant documents		strategy B: retrieved	strategy B: not retrieved	
strategy A	retrieved	n_{11}^r	n_{12}^r	
	not retrieved	n_{21}^r	n_{22}^r	
				nr

and consider the test

$$H_o: p_{12}^r = p_{21}^r \text{ vs } p_{12}^r \neq p_{21}^r.$$

Under the null hypothesis we have that

$$e_{11}^r = n_{11}^r,\ e_{12}^r = \frac{n_{12}^r + n_{21}^r}{2},\ e_{21}^r = \frac{n_{12}^r + n_{21}^r}{2}$$

and $e_{22}^r = n_{22}^r$, where e_{ij}^r are the expected values in the $(i,j)^{th}$ cell.

Under the null hypothesis we have 2 degrees of freedom, and 3 degrees of freedom under the alternative.

Therefore, using Wilks' likelihood ratio test we obtain that, if H_o is true, then

$$\sum_{i,j} \frac{(n_{ij}r - e_{ij}r)^2}{e_{ij}r} \sim \chi_1^2$$

That is

$$\frac{(n_{12}r - n_{21}r)^2}{n_{12}r + n_{21}r} \sim \chi_1^2$$

So now if $P(A > B) \neq P(n_{1.}r > n_{.1}r)$ but

$$P(A > B) = P(\frac{(n_{12}r - n_{21}r)^2}{n_{12}r + n_{21}r} > 3.84)$$

(rejecting equality at the 5% level).

Now recall of $A = \hat{p}_A = \frac{n_{11}r + n_{12}r}{nr}$ and recall of $B = \frac{n_{11}r + n_{21}r}{nr}$. So

$$P(A \quad B) = P(\frac{nr^2(\hat{p}_A - \hat{p}_B)^2}{n_{12}r + n_{21}r} > 3.84) > P(nr(\hat{p}_A - \hat{p}_B)^2 > 3.84) \quad .$$

Since we are interested in the case A > B, this is equivalent to

$$P(\hat{p}_A - \hat{p}_B > |\frac{3.84}{nr}|) \quad .$$

Putting $z = \frac{(\hat{p}_A - \hat{p}_B) - (p_A - p_B)}{\sqrt{\frac{p_A q_A + p_B q_B}{nr}}} \sim N(0,1)$

and assuming $p_A - p_B = 0.05$, one obtains that the probability is equal to

$$P(z > (7.68)^{\frac{1}{2}} - (0.05)\sqrt{2nr}) \quad .$$

So, although this argument is a better one for deciding if A is better than B for a particular request, and reduces to the chances of circular triads by allowing ties, the cost of this improvement is that the number of relevant documents required to be assessed increases. Also this method runs into problems if an attempt is made to apply it to precision, and so fallout should be used.

A final warning is that since the significance level is 5% (or 1%), if there are a lot of pairs to be compared the wrong conclusions can be drawn occasionally. The method of overcoming this is known as multiple comparison (see Duncan (1955)).

Overall,there is no very good way of providing for direct multi-strategy comparisons. However David's method can be used for indirect comparisons, via paired subsets, if more extended strategy evaluation is required. Since it is simply concerned with results of comparisons between pairs of strategies, these comparisons themselves may depend on either the Pool or Squares approaches.