Reprinted from: *Studies in Theory of Information Retrieval*. Budapest: Foundation for Information Science, 2007 (pp 9–22)

ON DOCUMENT POPULATIONS AND MEASURES OF IR EFFECTIVENESS

Stephen Robertson
Microsoft Research, 7 JJ Thomson Avenue, Cambridge CB3 0FB, UK ser@microsoft.com

Keywords: Search effectiveness, score distributions, sampling, document collection.

Abstract: Work on the statistical validity of experimental results in retrieval tests has concentrated on treating the topics as a sample from a population, but regarding the collection of documents as fixed. This paper raises the argument that we should also consider the documents as having been sampled from a population. It follows that we should regard a per-topic measurement as also having a per-topic noise or error associated with it, which may depend critically on the number of relevant documents for that topic. Some of the common measures used in retrieval testing are re-examined from this point of view. The examination is essentially theoretical, supported by limited simulation experiments.

1 Introduction

When we evaluate the effectiveness of information retrieval systems, it is usual to take measurements over a number of test queries or topics. In the usual laboratory scenario, exemplified by TREC, these measurements (for example Average Precision) are averaged over topics, and when significance tests are applied, the usual unit of measurement is the topic. For example, we might perform a pairwise test comparing two systems on each of the test topics, and treating each observation (that is, each per-topic measurement) as having been sampled from some distribution. The model here is clearly that the topics themselves are to be regarded as being sampled from a population of topics.

This view disregards the process by which we arrive at a measurement per topic. Such measurements are based on the collection of documents, which itself might be regarded as a sample from a notional population. There is a simple argument which suggests that in most such tests, the number of topics is a much more critical sampling issue than the number of documents (documents are plentiful and cheap, but topics and the associated relevance judgements are expensive). Thus tasks in TREC, for example, are typically evaluated on the basis of maybe 50 topics, but (at the least) many hundreds of thousands of documents. Probably because of these relativities, we have tended to ignore the issue of sampling of documents and whether this might have any bearing on the reliability of our conclusions, in favour of worrying about the topic sample.

However, this argument is misguided. Most of the measures in common use are based on relevance, and therefore depend at least in part on the number of relevant documents identifed for a topic. This number is likely to be relatively small – in tens or hundreds or possibly even single figures – for each topic.

We should then ask the question: does the reliability of the per-topic measurement that we make vary between topics? And in particular, does it depend on the number of relevant documents for that topic? It seems likely that it does, and if it does, then the measurements that we average across topics must be regarded as having a variable amount of noise. Thus we should probably place more reliance on an average precision value (say) derived from a topic with many relevant documents than on one with few. But *none* of the methods in common use takes this view.[1]

[1] One exception to this statement is the so-called micro-average method that can be applied to some measure, commonly reported in categorization experiments and used in the past in some retrieval experiments. It would be of interest to study this method following the ideas presented in the present paper; however, it is not obviously applicable to

Typically in evaluation we make use of measures (e.g. average precision) which are defined for a specific set of results from a specific collection of documents. How could we relate such a measure to an arbitrary (possibly infinite) *population* of documents? One possible answer would be to define a measure on the infinite population, such that the chosen test measure can be regarded as an *estimate* of the population measure. Just for example, uninterpolated average precision is defined by locating each actual relevant document in the actual ranked list, measuring precision (as a proportion) at that position, and averaging these values. What would be the general measure, definable for an infinite population, of which this would be an estimate?

Another possible answer is to approach the problem in the opposite direction, by considering increasing sample sizes. We then ask the question: if we take this measurement on samples of increasing size, do we approach a sensible limit as the sample size tends to infinity? We take these two versions of the question as equivalent at some level:

1. if there is a well-defined population measure of which the sample measure is a reasonable estimate, then the sample measure will converge to the population measure as sample size goes to infinity;
2. if the sample measure converges to a sensible limit as the sample size goes to infinity, this limit is a well-defined measure on an infinite population.

The primary question to be investigated in the present paper is: is it possible to interpret the measures in common use in this way? In cases where we can make such an interpretation, we may additionally be interested in how good the estimate is likely to be for smaller samples. Specifically, we may be interested in statistical precision, and in bias.

The differences between topics are a complicating matter in this investigation. One difference, as discussed, relates to the number of relevant documents (almost always small compared to the collection size, but nevertheless varying by several orders of magnitude). A second relates to the relative hardness of topics, a subject of much current interest (see e.g. (Voorhees, 2006)). These questions will be largely avoided in the present discussion, by the device of considering only a single topic at a time.

Some of the concerns of this paper are investigated by Cormack and Lynam (Cormack and Lynam, 2006). In particular, they consider the test collection of documents as a sample from some notional population, and investigate the statistical precision of estimates of MAP under these conditions, using a bootstrap method. They make a strong case for this sampling view of the document collection; they also argue that the combination of evidence from different topics should be treated in a different way, regarding each as in some sense a different experiment. However, they do not consider the question of whether specific measures *can* be interpreted in an infinite population, nor of possible biasing.

The present paper draws some material from Hawking and Robertson (Hawking and Robertson, 2003), much of which is devoted to studying subsampled collections of documents and the effect of subsampling on effectiveness measures. However, the assumption in the present paper is that we do not necessarily have access to the notional population from which we are sampling; indeed we might assume it to be infinite. The exploration in this paper is based on a combination of theory and simulation; no actual experimental results are reported here, although some connection is made with experiments reported by others.

Organization of the paper

In the following section, we discuss some ideas which are basic to the present approach, including the score-distribution model. In Section 3, we consider some of the commonly-used test measures, and attempt to interpret each as an estimate of a population measure. This leads us to the conclusion that some measures can be interpreted in this way and some cannot. For those that can, we also discuss estimation issues. In Section 4, we use simulation methods, in part to illustrate the estimation issues, and in part to confirm the difficulty of making suitable interpretations of some measures. Some related empirical evidence from previously reported real experiments is discussed in Section 5. In Section 6 we briefly discuss a class of topics for which one of the assumptions of the model is not appropriate. Finally, we draw some conclusions.

2 Basic notions

2.1 Document population and test collection

Considering a single topic at a time allows us to regard relevance as a primary variable in this investigation: we can assume that the (maybe infinite) population of documents is characterized in terms of this variable. More particularly, we make the usual assumptions about relevance (to this topic), namely that

ranked results.

it is a binary property of each document, defined independently of other documents. We now assume that we have sampled this population in order to provide the (finite) test collection.

We also assume that the ratio of relevant to non relevant for this topic is in some way determined from outside. (The proportion of the entire population that is relevant is traditionally known as *generality*, and will be denoted G below). If the entire population were assumed to be finite, then we can formulate this assumption in the following way: for a particular topic, the populations of relevant and non-relevant are finite and therefore the ratio of their sizes is well-defined. If the entire population is sampled without reference to relevance, then the expected generality in the sample is determined by this ratio. If, on the other hand, we assume the population is infinite, we have to abstract this notion. We assume instead that generality G is an inherent property of the topic (that is, that the probability that a random document from the population is relevant to the topic is a topic-dependent property).

The fixed-generality assumption is problematic for one class of queries: those for which it is reasonable to assume that there is only one relevant document (whatever the size of the collection). This issue is discussed in section 6 below.

In reality, of course, a single document collection is normally used for all topics. Although the collection looks different from the point of view of each topic (because of the different relevance conditions), it is actually made up out of the same documents; the document samples for each topic are not made independently. This is an extremely complex sampling situation, very hard to analyse; for the purpose of the present paper, we make the simplification that each topic's view of the collection may be treated independently.

2.2 Score distributions

We assume that given a topic, a system will assign a score to each document. As we have already assumed that the documents (identified as relevant or not) have been sampled from some notional large population – actually two populations, one for the relevant and one for the non-relevant documents – it follows that given a system, the scores themselves might be regarded as having been sampled, again from two distributions. These distributions (of scores of relevant and non-relevant documente respectively) form the basis of the arguments presented here. They depend on the specific system and the population of documents.

Actually this assumption is arguable. If the scoring system assigns scores to a document-topic pair without reference to any other documents, the statement is valid, but insofar as scores depend on other documents, the statement is suspect. In (Hawking and Robertson, 2003) this issue is discussed at length. Here the assumption will be made without further justification.

The notion of score distributions was introduced into IR by Swets (Swets, 1963), and has been developed in several more recent works (see e.g. (Robertson, 2007)). The usual issue concerns what specific distributional forms might be assumed and/or fitted to data. While certain specific distributional assumptions will be made for the simulation experiments below, this paper is concerned with arguments which are essentially independent of any such specific assumptions.

We note also that when documents are ranked by score, the ranking is invariant under any monotonic transformation of the score. In the sense that we are primarily interested in the ranks and not the scores, any arguments we choose to make about score distributions should themselves be invariant to such transformations.

The model is referred to here (as in (Hawking and Robertson, 2003)) as the SD (signal detection or score distribution) model. We denote the score distribution density functions of relevant and non-relevant documents as $f_R(s)$ and $f_N(s)$ respectively (s represents the score).

2.3 Multiple topics

The basic premise of this paper is that we look at a single topic at a time. However, the issue of averaging across topics (which is central to system evaluation) clearly needs consideration, and is indeed part of the motivation above for the present investigation.

The translation of the score distribution idea to multiple topics is not straightforward. The simplest way to do this would be to assume that the distributions of scores for relevant and non-relevant documents are the same for all topics.[2] However, we have much evidence to suggest that this is a bad assumption. One obvious reason for this is that scores seem very clearly to be not comparable across topics.

We could view this problem as one of normalising or calibrating scores so that they are comparable across topics (for example, making them into well-calibrated probabilities of relevance). However, this is a difficult problem in its own right, and would not

[2] In some sense, this assumption is implicit in the micro-averaging method mentioned in the previous footnote.

deal with the issue that (in our current understanding) some topics are harder than others. An alternative view is presented in the next section.

2.4 The big picture

We present here an overview of the conceptual model of sampling-and-measurement that informs the present paper. The paper only addresses a small part of this domain, but it is necessary to provide context.

We consider a single system and a test corpus, consisting of a set of topics and a collection of documents. Both topics and documents are assumed to have been sampled from large or infinite notional populations. We assume that (for this system) each topic has its own 'true' effectiveness (by some measure), which we would measure on the entire population of documents if we could; in the event, the best we can to is to make probabilistic inferences about it from the sample of documents that we have. The true effectiveness measure may depend, for example, on the hardness of the topic, and therefore can be expected to vary between topics.

So in order to generalize over topics, we must assume that the entire population of topics defines a distribution for this effectiveness measure (for the specific system), dependent on some (hyper-)parameters. The generalized effectiveness of the system will be represented by these hyperparameters. We would like to estimate, or generally make probabilistic inferences about, these hyperparameters, based in turn on the evidence about individual topic effectiveness.

This paper does not address the question of what this overall effectiveness distribution and its hyperparameters would look like; it focusses instead on the per-topic question. However, it is the belief of the present author that this second stage is necessary, and that the overview presented is consistent with the arguments made in (Cormack and Lynam, 2006). But in order to progress in this direction, it will be necessary to devise models which can take explicit account of both topic hardness and generality.

3 Effectiveness measures on infinite distributions

We suppose, then, that there exist for each topic arbitrarily large or infinite notional populations of both relevant and non-relevant documents, and that the observed documents represent samples from these populations. Furthermore, for a particular system and topic, these populations are expressed as score distributions. Now we should ask the following questions:

1. How would we measure effectiveness of the system on a topic if we *knew* the distributions in full?
2. Having defined a measure on the full distributions, how do we estimate it from the document samples that we have?
3. Can we interpret the usual per-topic measures used in IR (which are defined on the sample) as estimates of document-population measures?
4. (equivalently) Could we expect a per-topic measure defined on the sample to approach a reasonable limit as we increase sample size towards infinity?
5. How good are the sample estimates? Can they be improved?
6. If a traditional per-topic measure cannot be interpreted in these ways, what does this tell us about the measure?

We note again that we seek arguments that are invariant to monotonic transformations of the scores. In particular, virtually all measures used in IR are based on ranks, not scores, and are therefore invariant in this sense. If we do define a measure on the full distributions, we have to ensure that it has this invariance. Proving this invariance of a defined measure is fairly straightforward but tedious. Below, we simply state the invariance property for the measures to which it applies; the formal details are discussed in the Appendix. Furthermore, we also note that virtually all measures are 'top-heavy' in the sense of being heavily weighted to the top end of the ranking.

The above is a formidable series of questions, and the present paper will only scratch the surface of this space. A parallel set of questions will arise later, when we consider multiple topics; again, these are left for future work.

3.1 Recall and precision

Recall and precision are usually defined on retrieved sets rather than rankings of documents. If we define a retrieved set by explicitly thresholding a scoring function, then (under the above assumptions) we can easily define the population equivalent of these measures.

The recall equivalent is simply the probability that a random relevant document will be retrieved (i.e. that its score exceeds the threshold). It is a function of the relevant score distribution only. Furthermore, the usual measure recall (defined on the observed sample) is the obvious maximum likelihood estimate of this population measure. The population measure may be defined for a threshold t as:

$$F_R(t) = \int_t^\infty f_R(s)ds \qquad (1)$$

that is, as the cumulative distribution function calculated *from the right*. It is also useful to define the equivalent of *fallout*[3] as follows:

$$F_N(t) = \int_t^\infty f_N(s)ds. \quad (2)$$

The precision equivalent can be defined as the probability that a document scoring at or above the threshold is relevant; it is a function of both score distributions, looking like this:

$$H(t) = \frac{GF_R(t)}{GF_R(t) + (1-G)F_N(t)} \quad (3)$$

(this is exactly the usual formula relating recall, precision, fallout and generality, as reported for example in (Cleverdon et al., 1966); it can also be derived by simple probability manipulations from the above probability definitions of these parameters). In this case it is not so obvious that the usual sample proportion estimate is a good one. However, we may deduce that it is at least the maximum likelihood solution, from the observation that sampling randomly both relevant and non-relevant with the same probability, and selecting from both samples those whose scores exceed t, gives us a random sample of the correct subset of the combined population.

We note that if we assume that the distributions F_R and F_N are independent of generality G, then it follows from Equation 3 that precision is strongly dependent on G. This will have implications when we try to summarise evaluation data over multiple queries. We also note that the assumption would be a strong one; there may well be reasons why the parameters of the distributions are not themselves independent of generality. However, it is important to be clear that these other dependencies are extremely unlikely to cancel out the one already identified. Thus we must assume that the population parameter we are trying to estimate, H, is highly dependent on generality.

This raises an additional problem: that of estimating generality. We have assumed above that generality is a fixed property of a topic; however, the generality observed in a sample collection can only be an estimate of the true population generality. This compounds the issue of estimating measures which are in some way dependent on generality. The problem will be avoided in the simulation experiments reported below, by fixing generality in the sample rather than by allowing it to be determined by the sampling process. However, this is an unrealistic simplification.

[3] *Fallout* is a measure conventionally defined along with Recall and Precision, although not commonly used: it is the proportion of non-relevant documents that are retrieved. In other contexts it is known as the probability of a false positive.

All of the above population measures are invariant to monotonic transformations in the score variable (although of course the threshold t also has to be transformed).

3.2 Pairwise error probability

One measure which comes from the signal detection theory domain is the probability of a pairwise error, which may also be interpreted as the area under the ROC or Receiver Operating Characteristic curve – in the information retrieval context, this is the recall-fallout curve on linear scales. This measure is well-defined for the population, has the necessary invariance property, and can easily be estimated from a sample. The population definition may be expressed as:

$$\int_{-\infty}^{\infty} f_R(s)F_N(s)ds.$$

Pairwise error probability suffers as an IR effectiveness measure because it is not at all top-heavy. That is, it pays equal attention to pairwise errors way down the ranking as to those at the top. Thus a relevant document being lifted over 1000 non-relevants from rank 2000 to rank 1000 has 1000 times more effect on the measure than a relevant being lifted over one non-relevant from rank 2 to rank 1. Another difficulty is that of averaging over topics. The difficulty has to do with defining a single ROC curve which summarises a set of topics. The same problem arises in the context of the more common (in IR) recall-precision curves, and will be discussed in the following section. However, the general formulation of the pairwise error probability as defined here for a population (area under a curve = integral over some function of the distributions), and its invariance property, will inform some of the subsequent discussion.

3.3 Ranked results and the R-P curve

If we were to take points defined by the score threshold, and plot each point on the recall-precision graph, we could use the same arguments as above – the graph for the populations would be a well-defined function of the distributions, and the observed graph would be the obvious estimate of the population graph. This would work for a single topic. However, a recall-precision graph for a single topic is not usually regarded as very meaningful, and we look to averaging across topics for a system evaluation.

At this point, we run into all the problems mentioned in section 2.3. Thus we do not fix the points for merging across topics by score; this would involve bad assumptions about the compatibility of the score

distributions for different topics. The normal method is to use the recall level as the method of merging – in other words to measure precision at each level of recall, and average precision values across topics for a given recall level.

There is a question here about interpolation, since the number of documents relevant to each topic is different, and one cannot necessarily find exactly the recall level required for each topic. However, putting this question aside, there are two prior questions in the terms of the present paper: if we measure precision at a given recall level, is this a well-defined measure on the populations, and is our way of estimating it good?

The issue here is that instead of estimating recall at a given score threshold, we are estimating the score threshold to achieve a given recall. Suppose, for example, we want to measure precision at recall 10% for a given query. This would be well-defined in terms of populations (the recall level of 10% defined in terms of populations is just the 10th percentile of the relevant score distribution). This defines a score threshold, and we can use this to measure precision. However, we estimate this threshold as the score threshold that achieves 10% recall in the *sample*. It is not immediately clear how good an estimate this is, nor whether using this estimate in turn to estimate precision is good. In particular, it is likely that errors in the estimation of the threshold are likely to be magnified as errors in estimating the precision, because the latter will be highly affected by the density of non-relevant documents around the threshold.

Below, simulation evidence will indicate that for small numbers of relevant documents, this method is likely to give an over-estimate of the population measure (the smaller the relevant document sample the greater the bias). We suggest an explanation in the next section.

3.4 Average precision

Average precision (AP) may be regarded as the area under the recall-precision curve; it can also be thought of as a specific non-standard form of pairwise error probability, or rather the reverse ((Aslam et al., 2005) has a related interpretation). We can define it in a generative fashion which provides us with an equivalent measure on the population, as follows:

1. Choose a random relevant document d_1;
2. Choose a random document d_2 scoring at least as high as d_1;
3. Measure the probability that d_2 is relevant.

This definition indicates the connection with pairwise error; if the last step had said *non-relevant*, it would have been a form of pairwise error. However, the particular form of this definition provides the desired top-heaviness.

In the case of the sample, this definition translates readily as the usual non-interpolated definition of average precision: that is, average precision is the obvious estimate of this probability. In the case of the population, this definition is a simple integration over the score distributions:

$$J = \int_{-\infty}^{\infty} H(s) f_R(s) ds \tag{4}$$

This measure satisfies the invariance requirement.

However, if we ask whether the sample measure is a good estimate of the population measure, we run into the same problem as in the previous section. Furthermore, the formula gives us some clue as to why this might be. In the sample, we estimate the population f_R as being loaded onto the points representing the scores of the relevant documents; then we estimate precision H at those same points. Which means that we choose to estimate precision in the sample at exactly the points where this estimate peaks. If we had a different sample of relevants from which to estimate H than those we use for f_R, then the bias problem would not arise; but we use the same sample for the two purposes.

Another version of the same explanation is given in Section 4.3, grounded in what we actually do when calculating precision at the rank of each relevant document in the ranking.

We note also that J is almost certainly highly dependent on generality, because H has such a dependence.

Because we are dealing only with a single query, the same argument about average precision applies to the log of average precision, as used for GMAP (Robertson, 2006): $\log J$ is also a measure defined on the population which satisfies the invariance requirement – but also the log of (sample) average precision is likely to be a biased estimator of $\log J$.

3.5 Precision at rank n

The commonly-used measure precision at rank n (P@n, for example P@10) cannot be expressed as a measure on a distribution. This is consistent with the results discussed in (Hawking and Robertson, 2003), which confirm that in general in larger collections P@n will increase. P@n as it stands is not an estimate of any population parameter; under reasonable assumptions about the distributions in the SD model, it will tend to one as the sample size increases.

However, if we redefine the rank as a proportion of the total collection size, the measure can be inter-

preted in population terms. If the total collection size is N, we can consider the measure P@pN, where p is a (small) proportion: for example, if N is a million, and $p = 10^{-5}$, then P@pN is equivalent to P@10. This measure can now be defined in terms of the population as follows:

$$H(t) \text{ where } t \text{ satisfies } GF_R(t) + (1-G)F_N(t) = p \quad (5)$$

(this will be well-defined if the distributions f_R and f_N are continuous, but might have to be approximated if they are discrete). Again, this measure ($H(t)$ with this t) satisfies the invariance requirement: although the threshold t obviously gets transformed, the resulting $H(t)$ value is invariant.

Again, the question arises as to whether P@pN is a good estimate of $H(t)$. It seems to suffer from the same problem as precision at a specified recall level, that we have to estimate the appropriate threshold before estimating the measure. This would be done by choosing the document at rank pN – or (more likely) interpolating between two neighbouring documents if pN is not an integer. However, it is not so obviously biased, in the sense that the choice of threshold is not directly related to the sample distribution of relevants.

3.6 Some other measures

Success at rank n

This measure S@n (binary on a per-query basis, whether or not there are any relevant in the top n ranks) suffers from the same problem as P@n – under reasonable distributional assumptions, it will tend to one as sample size tends to infinity. The problem cannot be resolved by the method suggested above for P@n, however – under the same reasonable assumptions, S@pN will also tend to one. It seems that this measure makes no sense in the context of regarding the observed collection of documents as a sample from some large population.

Reciprocal rank

This measure is also hard to interpret in the present context (like the previous one, it in effect takes account of the first relevant document only). However, it was really defined for the situation where there is only one relevant document. That in itself seems incompatible with the notion of a potentially large population of relevant documents from which we have sampled. This issue is discussed further in section 6 below, where we see that reciprocal rank can be both interpreted and estimated in the one-relevant-only case.

NDCG

Normalised discounted cumulative gain, NDCG (Järvelin and Kekäläinen, 2000), is normally used for situations where there are multiple relevance grades, but it can be defined for the binary case as well.

It is another case of a measure for which there is no obvious parallel in the population. The explicit dependence on numerical ranks (the discount function) makes it hard to see how to make any equivalence, at least for any choice of discount function.

Truncated measures

In general any measure that is defined by truncating at a fixed rank, like P@n, will have problems. This applies to NDCG and AP, which are commonly truncated. In the case of NDCG@n, which is usually normalized by the maximum possible discounted cumulative gain at the same n, this will tend to one as sample size tends to infinity. In the case of AP, truncation implies assuming zero precision for relevant documents not yet retrieved; thus AP will tend to zero as sample size goes to infinity.

3.7 Discussion

Thus we seem to have rather few measures that are even analysable in these terms, and even they seem to have problems.

Is this an issue? If we regard the document collection as fixed, *and* persist in regarding each per-topic measurement as equally good, then no. The unit of measurement is the topic, and we observe the value we observe, and the only sampling issue is the sampling of topics.

However, the question raised above about the dependence of the measurement on the number of relevant documents disturbs the assumption about the equivalence of per-topic measurements, even if we continue to regard the collection as fixed. If we measure (say) NDCG, then the precision of each (per-topic) measurement must surely be affected by the number of relevant documents. We should have more confidence in a measurement based on 20 relevant documents than one based on just two. Then simply averaging the two allows the first (good) estimate to be polluted by the noise inherent in the second. The usual practice of averaging a measure over topics looks bad in these circumstance.

We may also be concerned with any possible bias in the estimate that might be dependent on the number of relevants.[4]

[4]Note that these are quite different questions from

It seems from the above arguments that we have no way of investigating these questions for NDCG or success at 10, say. We may be able to make such investigations for precision at fixed recall points or for average precision. The arguments above suggest that the usual estimates, as well as being less precise for fewer relevants, may also be biased. Both these questions are investigated in a very preliminary way in the simulation experiments that follow. The simulation experiments will also be used to illustrate the problem with NDCG.

4 Simulation experiments

The primary object of these simulations is to provide some evidence about the sampling behaviour of some of the measures discussed above, given varying sample sizes. In particular, we are interested both in the statistical precision of the estimates and in any biases that may be present. In addition, in the case of NDCG, we would like some insight into how the problem identified above might be reflected in the sampling distribution. The simulations are clearly very limited, and depend on the specific SD model assumptions, so are intended to be indicative only. Some of the conclusions are supported by other work with real experimental data – see the following section.

The starting point of the simulation is the previous work on specific continuous distributions (Robertson, 2007). There are several possible distributional assumptions (in pairs, one for relevant and one for non-relevant). Some have been shown to fit observed data well; there are also some theoretical considerations. Two models have been investigated for the present paper. The first has a Normal (Gaussian) distribution for relevant scores, and an exponential for non-relevant. This has been used by several authors because it fits observed distributions well (Arampatzis and van Hameren, 2001; Manmatha et al., 2001), but has a theoretical problem: it behaves unintuitively at either end of the score scale (Robertson, 2007). This behaviour is not normally problematic in the ranges of scores of practical interest, but the upper end could be important if we went to extremely large samples. The second model uses two Gaussian distributions of equal variance; this is less realistic but is not subject to these end effects.

All the results presented below (a small selection of the results obtained) are taken from the first model, avoiding the problematic regions. However, all the qualitative observations made about directions of change between samples of different size have been replicated in the second model. This may seem a little surprising (the exponential and Gausssian distributions are of very different shapes, and the recall-precision graph in the second case looks somewhat unrealistic). However, the conclusion fits well with the fact that we are primarily interested in properties that are invariant to monotonic transformations of the scores, which could change the shapes of distributions very drastically.

In order to simulate a collection with respect to a particular topic, using this distributional approach, we need to make the initial division between relevant and non-relevant documents. As indicated above, we assume that generality is a fixed property of the topic. The simplest way to instantiate this assumption is to sample in fixed ratio from the two distributions; a more realistic way would be to draw an initial binary variable with an appropriate probability, to determine relevance or non-relevance, and then to draw from the appropriate distribution. The simulations reported below use the simpler method. This approach avoids the problem about estimating generality from the sample; however, the problem will require investigation in the future.

The ideal way to use these distributions in the present argument would be to derive explicit formulae for the various document-population-based measures mentioned above, in terms of the parameters of the two distributions. For example, the measure J, the generalisation of average precision, is defined in Equation 4 and also makes use of Equation 3, and could in principle be evaluated as a function of the parameters of the normal and exponential pair of distributions. Then we could take samples of different sizes from these distributions and evaluate average precision under the usual definition, and compare the result with the theoretical value.

Unfortunately, these equations (for average precision at least) are intractable. Instead, we use the theoretical distributions to generate a single large sample, which we take as defining the population from which we will subsample. The value of average precision (usual definition) in the large sample is taken as the true value, and we consider to what extent small-sample results match the large-sample value. This is less satisfactory from the point of view of theoretical understanding; however, it does make the exper-

whether the measure correlates with or is affected by the generality. We are concerned not with whether 'real' NDCG is likely to be higher or lower for a topic with a higher density of relevant documents, just whether the estimate we get from the sample is likely to be more or less precise, and/or biased in either direction. However, for any multi-topic analysis, we will need to address both questions.

iments more compatible with other sampling experiments on real data (see Section 5 below). Also it reduces somewhat our dependence on the specific distributional assumptions: we only have to believe that the large sample is plausible, not that the distributions are truly exponential or normal.

4.1 Basic simulation

In the first simulation, we consider a fixed pair of distributions and fixed generality; we are concerned with both bias and error of estimation of various measures from a sample collection.

The exact form of sampling used below is as follows. The large sample consists of 12.8 million non-relevant scores, from an exponential distribution of mean 0.1. Most of these are thrown away; only the top-ranking scores are kept, enough to ensure that we always have at least 1000 top-ranking documents in any of the samples taken. 128 relevant scores are generated from a normal distribution of mean 1.0 and standard deviation 0.25. This large sample now becomes the population for subsequent sampling.

The population is now sampled at a rate $\frac{1}{2^n}$, where $n = 1,...,5$, following a procedure defined below. Each of these samples is repeated 10,000 times, and average precision is evaluated for every sample. We examine the mean and distribution of values for each n.

The procedure for the small samples is as follows. In order to control the number of relevant documents precisely (and in particular never have a sample with zero relevants), we take a $\frac{1}{2^n}$ sample *without replacement* from the 128 relevant documents. Thus when $n = 5$ we have exactly four relevant documents in each sample. The non-relevants, on the other hand, are sampled independently: we step down the ranked list and make a random draw for each non-rel to decide whether to include it or not. This makes use of the already-established ranking, which allows us to get away with throwing away most of the non-relevant scores before we start. The difference between the sampling methods for relevant and non-relevant may have had a very small effect on the results – see Simulation 2.

In Figure 1, we show a recall-precision graph. Note that this is for a single topic – it has not been smoothed, precision has been measured at every relevant document. The solid line represents the full large sample; this seems a perfectly reasonable recall-precision graph, a small confirmation that the distributional assumptions are at least reasonable. Each of the other lines represents a *single* sample from those indexed by n as above. Although there is inevitably some noise in these graphs, we already see some tendency to over-estimate precision at fixed recall levels, for smaller samples, as indicated in Section 3.3; we will see this more clearly below.

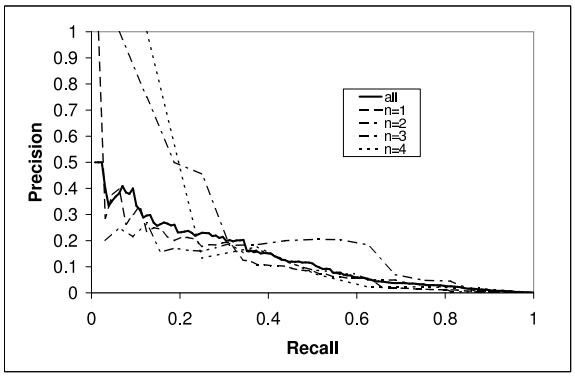

Figure 1: Recall-precision graph for the single topic in the whole population and in one each of the different sample sizes.

4.2 Simulation 1: AP

In order to calculate AP, we perform the usual calculation down to rank 1000. However, this misses a few relevant documents further down the ranking; the number may be affected by the sample size. In order to improve our estimates of true AP over the whole collection, we further estimate the small contribution of the missed relevants. We have their scores, and we can estimate how many non-relevants come above them in the ranking, from the distributional assumptions. This correction actually makes very little difference to AP in the range of samples considered, but is included for completeness (NDCG is more affected, see below.)

The results are shown in Table 1 and Figure 2. We see that (a) the standard deviation across samples of size $\frac{1}{2^n}$ increases with n (i.e. with decreasing sample size); and (b) the mean values have an increasing bias on the high side with increasing n. Thus in addition to getting less precise estimates from smaller samples, we get more bias.

Note on the error bars in Figure 2: These represent one standard deviation each side as given in the table. Thus many individual samples will fall outside this range. Note also that the precision of estimate of the mean over samples of a given size is much better: since each is based on 10,000 samples, the standard error of the mean is $\frac{1}{100}$ of the standard deviation, scarcely visible on the figure.

n	Rels	Mean AP	Std dev
All	128	0.138	
1	64	0.143	0.028
2	32	0.156	0.049
3	16	0.175	0.078
4	8	0.201	0.117
5	4	0.238	0.173

Table 1: Simulation 1: Average precision for simulated collection and samples of size $\frac{1}{2^n}$. Mean and standard deviation are over 10000 samples

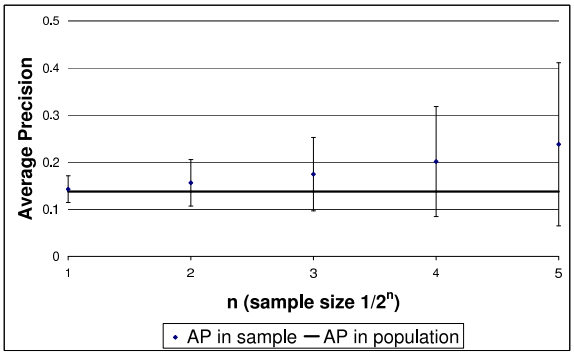

Figure 2: Simulation 1: Average precision for simulated collection and samples of size $\frac{1}{2^n}$. Error bars show a single standard deviation over samples each side.

4.3 Simulation 2: Precision

In this simulation, we investigate various ways of defining a point in the ranking at which to measure precision, for the simulation parameters already defined. We start with the traditional precision at fixed rank (P@n), with $n = 10$.

Table 2 shows how P@10 is reduced for smaller sample sizes. This fits with the discussion in Section 3.5 and with previously-reported results. The falling standard deviation in this case is probably an artifact of the falling mean (since precision cannot go below zero).

n	Mean P@10	Std dev
All	0.400	
1	0.362	0.113
2	0.322	0.128
3	0.258	0.123
4	0.213	0.113
5	0.152	0.092

Table 2: Simulation 2a: Precision at fixed rank for simulated collection and samples of size $\frac{1}{2^n}$.

Next we investigate the proposal of Section 3.5 denoted P@pN, where we fix the proportion of the collection rather than the absolute rank. Effectively we set p at 320/12.8m, but we define the threshold in two different ways. In the first, we set a score threshold and apply it to all samples; the score threshold is taken as the score of the 320th ranked document in the large collection. Thus the expected number of documents included in the calculation reduces from 160 at $n = 1$ to 10 at $n = 5$. However (taking $n = 5$ for example), this is not exactly the same as calculating precision at rank 10, because the exact number of documents exceeding this score may vary from sample to sample. In the second method, we measure precision at exactly the intended rank (from 160 down to 10); this is more comparable to what would be done in a real test. Results are given in Table 3.

n	Rank	Mean Prec	Std dev	Mean Prec	Std dev
		Method 1		Method 2	
All	320	0.150			
1	160	0.150	0.017	0.152	0.016
2	80	0.151	0.029	0.152	0.028
3	40	0.151	0.045	0.152	0.043
4	20	0.153	0.068	0.152	0.064
5	10	0.154	0.104	0.152	0.092

Table 3: Simulation 2b: Precision at rank for simulated collection and samples of size $\frac{1}{2^n}$. Method 1 calculates precision at the threshold given by rank 320 in the collection, Method 2 at the given rank in the sample.

For Method 1, we would expect the estimated precision values to be the same (within sampling error) for all samples. In fact, there is a very slight increase as we reduce sample size. This is probably the effect of the different sampling methods for relevant and non-relevant items, mentioned above. Method 2 is remarkably stable for smaller samples, perhaps slightly more so than Method 1.

Finally, we evaluate precision at a given recall level, set at 50%. Again, we calculate the precision value in two different ways. In the first, we set as score threshold the score of the 64th ranked relevant document in the large collection. Thus the expected recall level is 50%, though the actual recall level at this threshold will vary between samples. In the second method, we measure precision at the 50% recall level in the sample, at the rank of relevant document number 32 ($n = 1$) to 2 ($n = 5$); this is more comparable to what is normally done in tests. Results are given in Table 4.

Method 1 shows exactly the same pattern as for

n	Rank	Mean Prec	Std dev	Mean Prec	Std dev
		Method 1		Method 2	
All	64	0.116			
1	32	0.112	0.010	0.109	0.028
2	16	0.113	0.017	0.118	0.053
3	8	0.113	0.027	0.140	0.092
4	4	0.114	0.040	0.186	0.156
5	2	0.115	0.060	0.274	0.260

Table 4: Simulation 2c: Precision at recall level for simulated collection and samples of size $\frac{1}{2^n}$. Method 1 calculates precision at the threshold given by recall=50% in the collection, Method 2 at recall=50% in the sample.

P@pN; this is not surprising, since both amount to fixing the same threshold score value for all samples. Method 2, however, shows very strongly the kind of effect seen for average precision, that there is increasing upward bias for smaller samples.

We reiterate here, in a somewhat more concrete form, the reason for this upward bias suggested in Section 3.4. For a given topic and ranked list of results, if we perform no clever interpolation, the actual shape of the recall-precision graph is a series of points in vertical columns. This may be seen as follows. Suppose we have R total relevant documents; then the observed recall values will be $\{0, \frac{1}{R}, \ldots, \frac{R-1}{R}, 1\}$. Stepping down the ranked list, when we encounter a relevant item, we fix a point on the graph at one of these x-values. Thereafter, any succeeding non-relevant (until the next relevant) represents the same recall but a lower precision. Now, with Method 2 (which simulates what we really do for uninterpolated average precision), we choose to measure precision *at the highest point in each column*. The fewer the number of relevant documents, the fewer the columns, and the greater this column effect.

This analysis might suggest various possible solutions. We might for example perform some averaging or smoothing over the points in a column. However, it is hard to see how this could be done consistently, since for example the last column will contain an extremely large number of points, almost all of them microscopically close to zero. Another suggestion might be some form of interpolation; however, the traditional forms of interpolation on the R-P graph (being based on the usual inverse relationship between recall and precision) probably all have the same upward bias. It is in fact quite hard to see how to devise an unbiased method.

4.4 Simulation 3: Generality effect

The next simulation is based on fixed distributions but variable generality; we confirm that under these conditions, a measure like average precision is indeed affected by generality.

This time we construct the large sample with 3.2 million non-relevant and 128 relevant; we keep the number of non-relevant fixed, but sample the relevant on the same basis as previously. Thus different n values give different generality – $n = 2$ gives the same generality as in the other simulations. As noted above, measures such as precision or average precision are likely to be highly dependent on generality (decreasing as generality decreases). But this dependence is compounded with the biases revealed in the previous simulation, based on sample size. Thus the decrease in mean average precision as n increases in Table 5 is a combination of the decrease resulting from the reduced generality and the increase due to the small-sample bias; but we see clearly that the former effect is much stronger than the latter. Also the variance of average precision over the samples increases.

n	Rels	Mean AP	Std dev
All	128	0.304	
1	64	0.208	0.032
2	32	0.141	0.044
3	16	0.099	0.053
4	8	0.076	0.067
5	4	0.063	0.087

Table 5: Simulation 3: Average precision for a fixed collection size, but relevant document samples of size $\frac{1}{2^n}$.

We note again that the assumption made here that the population parameters themselves are independent of generality is a strong one, for which we have no real justification.

4.5 Simulation 4: NDCG

The final simulation evaluates NDCG – the discount function is taken as $\log(\text{rank} + 1)$, and since there is only one level of relevance other than non-relevant, the gain is treated as one. As with AP, we evaluate DCG exactly for each sample down to rank 1000; we then estimate the remainder using the method described for AP. Note that NDCG is affected by this tail much more than AP; the log-based discount function causes a much thicker tail. For reasons discussed in Section 3.6, we do not evaluate NDCG with a fixed truncation point.

Results are shown in Table 6. Here we see (a) that the precision of estimation of NDCG from samples does indeed fall with falling sample size, and (b) that the mean value decreases as the sample size decreases. The differences from the true population mean are quite large (all the differences between mean NDCG values are significant: for example, in the case of $n = 5$, the 95% confidence interval for the true mean over samples of this size is approximately (0.486,0.494)).

n	Rels	Mean NDCG	Std dev
Collection	128	0.639	
1	64	0.602	0.034
2	32	0.572	0.059
3	16	0.541	0.090
4	8	0.514	0.126
5	4	0.490	0.172

Table 6: Simulation 4: NDCG for simulated collection and samples of size $\frac{1}{2^n}$.

It is very clear that NDCG would increases with increasing sample size. The pattern suggests that NDCG would tend to one as the sample size tends to infinity; however, we have not demonstrated this.

5 Comparison with previous work

In (Soboroff, 2004), it is shown that several currently-used measures are particularly unstable when applied to topics with very few relevant documents. This result fits very well with the arguments presented above about the statistical precision of estimates.

In (Hawking and Robertson, 2003), results are presented of mean average precision on samples of different size from a large collection. These are means over 50 topics, rather than for a single topic, but on the basis of the above simulation results, we would expect to see that the sample results over-estimate the full-collection results, and the effect should increase for smaller samples (the extent of the over-estimation will vary by topic, but all the biases should be in the same direction).

What is observed in Figure 19 in that paper is that there appears to be a small rise in the sample estimates as we reduce the sample size, down to a 20% sample. But for 10% the movement seems to reverse. However, there is a confounding factor in these results. The evaluation was done with the trec_eval program, which excludes any topics with no relevant documents. Obviously the topics with the fewest relevant documents to start with (which are the ones most subject to the expected effect) are the most likely to be excluded, and are more likely to be excluded from the smaller sample results. In fact the 10% samples were significantly affected by this rule, and the results for this point are therefore based on significantly different topic sets than the larger samples.

In (Yilmaz and Aslam, 2006), a similar experiment is reported, with various sample sizes, but with the rule that if a sample produces a set with no relevant documents, then it is discarded and another sampling is made. Figure 4 in that paper shows a consistent over-estimate of mean average precision from the sample.

6 Single-relevant-item searches

For some classes of topics, it is a reasonable assumption that there exists only one relevant item in the database. Known-item searches, and web searches for home pages or named pages, come into this category. Such topics sit uneasily with the assumption of fixed generality, that if we draw a larger sample of documents as the full collection, we will maintain the proportion or density of relevant items by getting more of them.

We can think of this situation as one where we still draw the relevant document from a large population of hypothetical relevant documents, but having drawn one, we are constrained not to draw any more. While the assumption of a population of hypothetical relevant documents may seem a little strange for this situation, we may think of it as all the ways the author of the page *might* have chosen to write it. But having written it once, it does not make sense to write it again.

The commonest measure for this situation is the reciprocal rank. This might be interpreted in two ways as a population measure. First, the reciprocal rank is identical to the average precision in the case of a single relevant document; therefore it could be interpreted in the same way as average precision. But since we are assuming a single relevant document, this is a case of minimal generality and therefore maximal bias as identified above. A second interpretation is as the inverse of the pairwise error probability (taking the inverse provides the desired top-heaviness). The maximum likelihood estimate of the inverse pairwise error is

$$\frac{N}{r-1}$$

where N is the collection size and r is the rank of the single relevant document. Unfortunately this is unde-

fined if $r = 1$, a very common situation. Replacing it with $\frac{N}{r}$ looks like a simple smoothing to avoid this problem; and replacing this again with $\frac{1}{r}$ is a simple normalisation to get it into a reasonable range (for a fixed collection). It is likely that one could justify a better form of smoothing, and thus make a strong interpretation of reciprocal rank.

7 Final discussion

Over the last few years, considerable effort has been devoted to questions of statistical variation, or error, or significance in the results of retrieval tests. In this paper, we argue that this work has concentrated on just one of the two possible sources of statistical variation or error. That is, it has been assumed that the only sampling process involved is the sampling of topics or queries.

This paper has presented a very preliminary exploration of the consequences of considering the other source of statistical variation, namely the sampling of documents. These consequences are quite complex, and suggest firstly that only a few of the measures commonly used in IR experiments can be analysed in these terms at all; the others, however good they may be as pragmatic measures of effectiveness on a given collection, are not capable of generalisation. Secondly, the analysis suggests that even for those measures that can be analysed in these terms, there are serious issues of estimation which present methods do not address.

Again, we have not addressed directly the issue of multiple topics. Nevertheless, in the light of the analysis presented in this paper, the normal practice of averaging measures across topics (e.g. MAP, mean average precision in the usual sense) looks increasingly fragile. First we have shown that the expected error in estimation for a given topic depends very heavily on generality, so that our knowledge of the true average precision value is simply less good for a topic with few relevant documents. Secondly we have shown (for average precision) a systematic bias, also based on generality. Thirdly, the true measure itself is also likely to be highly correlated with generality. Averaging over topics of differing generality looks highly suspicious under these conditions.

This paper only scratches the surface of the problem. It is intended to open up the subject to further debate.

ACKNOWLEDGEMENTS

Thanks to Dave Hawking and Emine Yilmaz, as well as to my colleagues at MSR, for discussions on these issues.

REFERENCES

Arampatzis, A. and van Hameren, A. (2001). The score-distributional threshold optimization for adaptive binary classification tasks. In (Croft et al., 2001), pages 285–293.

Aslam, J. A., Yilmaz, E., and Pavlu, V. (2005). Expected average precision. In Marchionini, G., Moffat, A., Tait, J., Baeza-Yates, R., and Ziviani, N., editors, *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34, New York. ACM Press.

cikm06 (2006). *CIKM 2006: Proceedings of the 13th ACM Conference on Information and Knowledge Management*, New York. ACM Press.

Cleverdon, C. W., Mills, J., and Keen, E. M. (1966). *Factors determining the performance of indexing systems*. College of Aeronautics, Cranfield, U.K. (2 vols.) Aslib Cranfield Research Project.

Cormack, G. V. and Lynam, T. R. (2006). Statistical precision of information retrieval evaluation. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–540, New York. ACM Press.

Croft, W. B., Harper, D. J., Kraft, D. H., and Zobel, J., editors (2001). *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York. ACM Press.

Hawking, D. and Robertson, S. (2003). On collection size and retrieval effectiveness. *Information Retrieval*, 6:99–150.

Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In Belkin, N. J., Ingwersen, P., and Leong, M.-K., editors, *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, New York. ACM Press.

Manmatha, R., Rath, T., and Feng, F. (2001). Modelling score distributions for combining the outputs of search engines. In (Croft et al., 2001), pages 267–275.

Robertson, S. (2006). On GMAP – and other transformations. In (cikm06, 2006), pages 78–83.

Robertson, S. (2007). On score distributions and relevance. In Amati, G., Carpineto, C., and Romano, G., editors, *Advances in Information Retrieval: 29th European Conference on IR Research, ECIR2007*, pages 40–51, Berlin. Springer.

Soboroff, I. (2004). On evaluating web search with very few relevant documents. In Järvelin, K., Allan, J., Bruza, P., and Sanderson, M., editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 530–531, New York. ACM Press.

Swets, J. A. (1963). Information retrieval systems. *Science*, 141(3577):245–250.

Voorhees, E. M. (2006). Overview of the TREC 2005 robust retrieval track. In Voorhees, E. M. and Buckland, L. P., editors, *The Fourteenth Text REtrieval Conference, TREC 2005*. Gaithersburg, MD: NIST. `http://trec.nist.gov/pubs/trec14/t14_proceedings.html`.

Yilmaz, E. and Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgements. In (cikm06, 2006), pages 102–111.

APPENDIX

The invariance requirement

In section 3, we stated the requirement that measures defined on the population distributions of scores should be invariant under monotonic transformations of the scores, because such transformations do not affect the rank order of the documents. Here we sketch a formal proof for some such measures that they do indeed satisfy this requirement.

We suppose a monotonic strictly increasing tranformation ϕ on the score scale. We denote the original and transformed scores by s and s' ($s' = \phi(s)$), and similarly the original and transformed thresholds on the scores by t and t'. We consider how the distribution has to be transformed. It follows from monotocity that $P(s > t) = P(s' > t')$; hence any function or measure defined as a cumulative distribution transforms directly: $F(t) = F(t')$. Since the density function $f(s)$ is the derivative of the cumulative distribution, it follows that

$$f(s) = f(s')\frac{ds'}{ds} = f(s')\frac{d\phi(s)}{ds}$$

(since $\frac{d\phi(s)}{ds}$ is always positive, and the two functions integrate to the same value over the full range, this equation defines a valid transformation between distributions).

The measures corresponding to recall and fallout defined at a threshold ($F_R(t)$ and $F_N(t)$) are therefore invariant, as is the measure corresponding to precision at a threshold, $H(t)$, provided that the threshold itself is suitably transformed. In order to demonstrate invariance for other measures such as J (equivalent to average precision) or pairwise error, we simply equate within each integral the component $f(s)ds$ with the component $f(s')ds'$. In both these cases the integral is over the full range $(-\infty, \infty)$, thus no transformation of the bounds is required (if either score range is actually bounded, we simply assume without loss of generality that the density is zero outside the bounds).

We note also that any measure that assumes and makes use of the interval property of scores will not satisfy the invariance requirement. Specifically, for example, anything based on the means or other obvious properties of the score distributions will fail this test. The mean in general *does not* survive a monotonic but non-linear transformation.