Hypothesis Testing with Incomplete Relevance Judgments

Ben Carterette and Mark D. Smucker
{carteret, smucker}@cs.umass.edu
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003

ABSTRACT

Information retrieval experimentation generally proceeds in a cycle of development, evaluation, and hypothesis testing. Ideally, the evaluation and testing phases should be short and easy, so as to maximize the amount of time spent in development. There has been recent work on reducing the amount of assessor effort needed to evaluate retrieval systems, but it has not, for the most part, investigated the effects of these methods on tests of significance. In this work, we explore in detail the effects of reduced sets of judgments on the sign test. We demonstrate both analytically and empirically the relationship between the power of the test, the number of topics evaluated, and the number of judgments available. Using these relationships, we can determine the number of topics and judgments needed for the least-cost but highest-confidence significance evaluation. Specifically, testing pairwise significance over 192 topics with fewer than 5 judgments for each is as good as testing significance over 25 topics with an average of 166 judgments for each—85% less effort producing no additional errors.

Categories and Subject Descriptors: H.3 Information Storage and Retrieval; H.3.4 Systems and Software: Performance Evaluation

General Terms: Experimentation, Measurement

Keywords: information retrieval, evaluation, hypothesis testing, test collections

1. INTRODUCTION

Much work on retrieval systems is incremental: small changes to existing algorithms creating small gains in performance. Over time, small gains can build to substantial improvements. But small performance changes can happen for no reason but random chance, and whether they're worth pursuing further cannot be evaluated by visually inspecting retrieval results. We need statistical *hypothesis tests* to instruct us on whether a small change is worth following up on, or whether a line of research should be dropped.

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

In IR, hypothesis tests are performed over a set of queries, which are input to a system to produce a ranked list of documents. Each ranked list is evaluated against a set of relevance judgments that indicate whether each document is relevant to the query. Unless a large set of relevance judgments is already available, they must be acquired by having human assessors read and judge documents. This is a very time-consuming process, and as a result, there has been a great deal of recent interest in small sets of judgments. But an evaluation over a small set of judgments will produce errorful measures of performance; it seems clear that they must affect the conclusions drawn from a hypothesis test.

We treat a hypothesis test as a binary decision-maker: the null hypothesis is either rejected or not rejected. For the decision to have any meaning, it must be tied to some implication about the reason for it. If the null hypothesis is rejected, we want it to be because we are unlikely to have observed a particular sample in a world in which that hypothesis is true. This is captured by the *accuracy* of the test: a test with high accuracy is not likely to falsely reject the null hypothesis.

On the other hand, if the null hypothesis is *not* rejected, we want it to be because our sample is unlikely to have been observed when the null hypothesis is *not* true. This is captured by the *power* of the test: a test with high power is likely to reject the null hypothesis when it is false. Power is important but subtle. If we decide to drop a line of research because it did not produce a significant result, we must be certain that the power of the test is high. If it isn't, then the failure to reject the null hypothesis is not meaningful.

Our goal in this work is to investigate how incomplete relevance judgments affect the conclusions we draw from hypothesis tests. Our focus is on power; as we will see, incomplete relevance judgments, when uncertainty is properly accounted for, do not affect the accuracy of the test.

We begin with a brief look at previous work on hypothesis testing in information retrieval. We then provide a tutorial on the sign test with special emphasis on the notion of power. This leads into our first major result: an expression to determine how many topics should be used to maintain power when there is uncertainty due to relevance judgments. After that, we describe how to estimate the uncertainty due to relevance judgments, leading to our next major result: a model for estimating the number of judgments needed to reach a given level of uncertainty with a given number of topics. We can then define a cost function for experimentation to find the optimal number of topics and judgments needed to run a significance test that has high power.

2. HYPOTHESIS TESTING IN INFORMATION RETRIEVAL

Investigations into the appropriate hypothesis tests to use in information retrieval experimentation go back at least as far as van Rijsbergen's classic 1979 textbook [10]. Van Rijsbergen discusses the sign test, Wilcoxon sign rank test, and t-test, and concludes that since little is known about the distribution of evaluation measures, only weak tests like the sign test can be used.

Zobel [14] and Sanderson & Zobel [8] undertook an empirical investigation of hypothesis test performance on retrieval systems that had been submitted to TRECs (Text REtrieval Conference) over the years. As these systems represent real retrieval systems over real topics that people might be interested in, they provide an opportunity to evaluate and compare tests on real data. Both works also investigate the effect of reducing assessor effort on evaluation by using other evaluation measures or reduced-depth pools of judgments.

Recently, there has been some interest in whether small test collections can generalize. There are two notions of "generalization" in retrieval experimentation: generalization to a new set of systems that did not contribute any judgments to the set, and generalization to new topics that have not been seen before. The latter is the domain of hypothesis testing. Recent work on the TREC Web and Terabyte tracks has suggested that more topics and fewer relevance judgments provide evaluations as good as a few topics with a lot of relevance judgments.

Cormack & Lyman investigated the effect of small test collections on the power of a test empirically, concluding that good evaluation can be provided by many topics with a small number of judgments for each [3].

The work most closely related to this one is Jensen's Ph.D. thesis [6]. He undertook a careful empirical investigation into the power of hypothesis tests over large sets of topics evaluated on only the top retrieved results, additionally investigating the effect that automatically-assigned relevance judgments have on power. His two findings are that large topic sets are necessary when evaluating over few retrieved results, and that automatic relevance assignments decrease power.

Our conclusions are the same as the previous two works: more topics with fewer judgments is at least as good as full sets of judgments. Above that, our contributions are an analytic investigation of the sign test leading to a cost function for determining the optimal number of topics and judgments needed, and an empirical evaluation of that cost function on real IR systems. We have elected to focus on the sign test due to its simplicity; we hope to perform a similar analysis for the t-test.

3. THE SIGN TEST

The sign test is appealing as it is one of the easiest to implement, the easiest to understand, and makes the fewest assumptions about the data. For simplicity, we will focus on the one-sided sign test; our results can be extended without much difficulty to the two-sided test.

The two hypotheses in the one-sided sign test are:

$$H_0: \quad \theta \leq \theta_0$$
$$H_1: \quad \theta > \theta_0$$

We assume we have n i.i.d. Bernoulli trials $Y_1, Y_2, ..., Y_n$, each of which having probability of success θ, i.e. $P(Y_i = 1) = \theta$. The test statistic is $S = \sum_{i=1}^{n} Y_i$, the number of successes. S has a binomial distribution $Binom(n, \theta)$. If the null hypothesis is true, the maximum expected number of successes is $n\theta_0$, with a maximum variance of $n\theta_0(1-p\theta_0)$. If S is much greater than that expectation, it is unlikely that the data is distributed according to the null hypothesis. θ_0 is generally chosen to reflect the hypothesis that each trial is equally likely to be a success or failure, $\theta_0 = \frac{1}{2}$. If S is unlikely to have occurred when $\theta_0 = \frac{1}{2}$, then we may reasonably conclude that the observed values did not occur by chance.

For a given level of significance α, there is at least one "critical value" c_α such that $P(S \geq c_\alpha | n, \theta_0) = \sum_{i=c_\alpha}^{n} \binom{n}{i} \theta_0^i (1-\theta_0)^{n-i} < \alpha$. If the observed S is greater than the maximum c_α, we may reject the null hypothesis as being unlikely.

In this formalism, α is the expected accuracy of the test. It defines the probability of making a Type I error, or false positive, of rejecting the null hypothesis when it is not true. Figure 1(a) shows the distribution of S under the null hypothesis for $n = 50$; the shaded region is the probability of rejecting the null hypothesis when $\alpha = 0.05$. If H_0 is true, the area of the shaded region corresponds to the probability of making a Type I error.

As n increases, the binomial distribution converges to a normal distribution:

$$\frac{S - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \rightsquigarrow N(0,1)$$

Therefore for large n, we can use a normal distribution function to approximate the binomial, avoiding the computational difficulty of calculating $\binom{n}{i}$. The normal cumulative density function with zero mean and unit variance is generally denoted by Greek letter Φ; $\Phi(x) = P(X < x)$ is the probability that normalized random variable X takes on a value less than x. Φ is defined as the lower tail of the normal distribution, but since our alternative hypothesis is that $\theta > \theta_0$, we need the upper tail.

$$P(S \geq c_\alpha | n, \theta_0) \approx 1 - \Phi\left(\frac{c_\alpha - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\right)$$

We can estimate c_α using the normal quantile function Φ^{-1}.

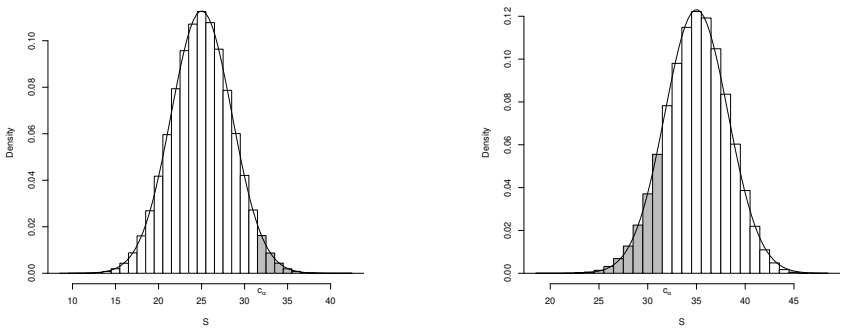
$$Z_\alpha = \frac{c_\alpha - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \approx -\Phi^{-1}(\alpha)$$

The normal approximation is generally acceptable for $n > 25$ [13].

3.1 Power

The complement to accuracy is *power*. Power reflects the probability of making a false negative error, that is, failing to reject the null hypothesis when it is false. This is also known as Type II error and usually denoted β. Power is $1 - \beta$, the probability that the null hypothesis will be rejected when it is false.

Power is relevant when the null hypothesis is false; therefore we need $\theta > \theta_0$. It will also be useful to think in terms of a population "effect size" $h = \frac{\theta - \theta_0}{\theta_0}$ [2]. This is the percent increase in successes above what would be expected by the null hypothesis. If the null hypothesis is true, then effect size $h = 0$. For the purposes of analyzing the power of the one-sided sign test, we will define $h > 0$.



(a) Null hypothesis true.

(b) Null hypothesis false.

Figure 1: Distribution of S depends on θ. On the left, the gray area represents the probability of making a Type I error. On the right, the gray area is the probability of making a Type II error. The error region is bounded by the critical value c_α in both.

For a given significance level α and sample size n, the power of the one-sided sign test is as follows. Let c_α be the maximum critical value at which we will reject the null hypothesis. For a given population success rate θ, power is defined as:

$$\begin{aligned} 1-\beta &= P(S \geq c_\alpha | n, \theta) \\ &= \sum_{i=c_\alpha}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} \end{aligned}$$

Figure 1(b) shows the region of the binomial distribution that would produce Type II errors when $\theta = 0.7$.

From this equation we can see how each variable in the test (n, α, h) affects the power. Increasing sample size n increases power. Increasing significance level α increases c_α and therefore decreases power. Increasing effect size h entails an increase in true success proportion θ, which increases power. Figures 2(a) and 2(b) show how power is affected as effect size, sample size, and significance level α change.

In reality, we can control n and α, but we have no control over h. In determining the number of trials, then, we must consider the minimum effect size we would like to be able to detect with high probability, while keeping the probability of making a false positive (Type I error) low.

We can also express power in terms of the normal approximation. When $\theta_0 = \frac{1}{2}$,

$$\begin{aligned} 1-\beta &\approx 1 - \Phi\left(\frac{c_\alpha - n\theta}{\sqrt{n\theta(1-\theta)}}\right) \\ &= 1 - \Phi(Z_\alpha - h\sqrt{n}) \\ &\approx \Phi(\Phi^{-1}(\alpha) + h\sqrt{n}) \end{aligned}$$

where Φ is the normal density function with zero mean and unit variance. (For details on the derivation of this expression, see Cohen [2].)

We have introduced a lot of notation up to this point, and there is still more to be introduced. Table 1 provides an easy reference to the notation and its meaning.

3.2 Sign Test Example

Suppose we have two retrieval algorithms, A and B, and a sample of $n = 50$ topics. Our null hypothesis is that algorithm A outperforms algorithm B half the time on the population of topics, i.e. whether one is better than the other is essentially random.

$$\begin{aligned} H_0 &: \quad \theta \leq \tfrac{1}{2} \\ H_1 &: \quad \theta > \tfrac{1}{2} \end{aligned}$$

For $n = 50$ and $\alpha = 0.05$, the critical value c_α is 32: if A outperforms B on at least 32 of the 50 topics, we will reject the null hypothesis.

Suppose we know that $h = 0.4$, i.e. A outperforms B on 70% of the topics in the population.[1] Using the normal approximation, the power of a test with 50 topics and $\alpha =$

[1]Obviously we have no practical way of knowing this, but the assumption will help demonstrate power.

sign test notation

θ_0	null hypothesis about proportion of successes.
θ	population proportion of successes.
n	sample size.
Y_i	Bernoulli random trial with $p(Y_i = 1) = \theta$.
S	observed success count. $S = \sum_{i=1}^{n} Y_i \sim Binom(n, \theta)$.
α	significance level; probability of Type I error.
c_α	critical value for significance level α.
Z_α	normalized critical value.
β	probability of Type II error; power $= 1 - \beta$.
h	effect size $h = \frac{\theta - \theta_0}{\theta_0}$.

uncertainty notation

$\widehat{Y}_i$	estimated Y_i in the presence of uncertainty.
λ	certainty $\lambda = P(\widehat{Y}_i \mid Y_i)$.
h'	adjusted effect size based on certainty λ.
n'	adjusted sample size based on certainty λ.

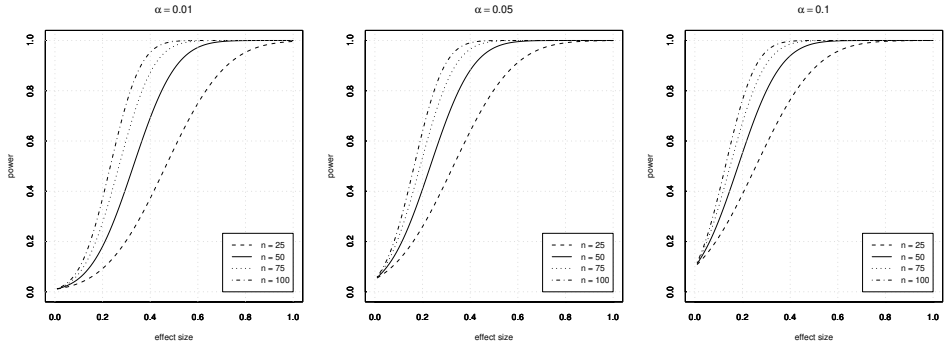
evaluation notation

m	number of documents in the collection.
X_i	Bernoulli random trial for relevance of a document.
ΔMAP	difference in mean average precision.
$\hat{j}(\lambda, n)$	estimated number of judgments to reach λ.
γ	coefficients in $\hat{j}(\lambda, n) = \exp(\gamma_0 + \gamma_1 \log \lambda + \gamma_2 \log n)$.

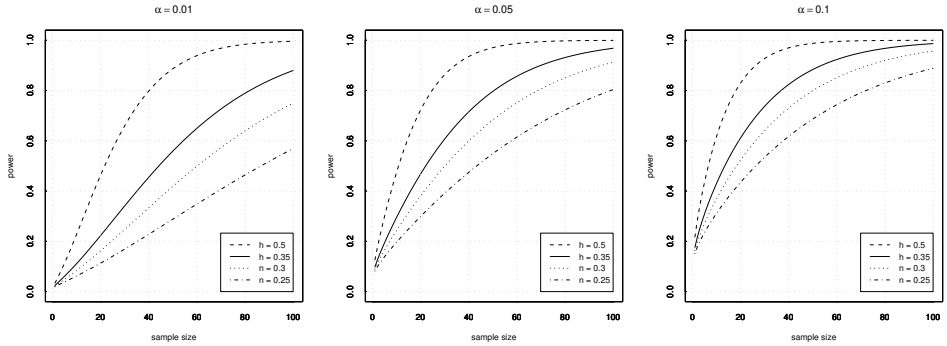
cost notation

C_j, C_t	cost of making a judgment and developing a topic.
$C(\lambda, n)$	total cost incurred to experiment with uncertainty λ and original sample size n.

Table 1: Table of symbols.



(a) Increasing effect size increases power.



(b) Increasing sample size increases power.

Figure 2: Effect of sample size, effect size, and significance level on power.

0.05 is about 0.882, so the probability that we will draw a sample of 50 topics and fail to reject the null hypothesis on the basis of that sample is about 0.12.

Since 50 topics has become the standard for IR experimentation, it is interesting to calculate the power of the sign test to detect varying effect sizes with $n = 50$. If we want 80% power and 95% accuracy, the effect size must be at least 0.35, i.e. the better system needs to be better on at least 68% of the topics. For 60% power, the effect size must be at least 0.25, or 62.5% of the topics. For 95% power, the effect size would have to be 0.47; A would have to outperform B on 73.5% of topics.

3.2.1 Evaluating Power

Does the theory pan out? Are we able to detect significant differences between retrieval systems at the rate predicted by the analysis above?

In IR hypothesis testing studies, topics are generally taken to be i.i.d. samples from some population, and retrieval runs that were submitted to TREC tracks are used to test hypotheses about hypothesis tests. We will follow this approach, using the 249 topics from the 2004 TREC Robust track and the 110 submitted systems [12].

We will treat the 249 topics as a population from which we sample uniformly at random. We will take the population effect size h to be the effect size over the 249 topics in the "population".

We randomly selected samples of n topics from the set of 249. We performed sign tests on all 5995 pairs of systems. As stated above, with $n = 50$ we can detect an effect size of $h = 0.35$ with power 0.8 at significance level 0.05 (using the normal approximation). If the analysis is correct, we should see that for all pairs with a population effect size of $h \approx 0.35$, we correctly rejected the null hypothesis for 80% of them. We will refer to the percentage for which we do reject the null hypothesis as the "observed power". This phrase is sometimes used to mean "post-hoc power"; the concept of post-hoc power has been discredited by Hoenig & Heisey [5]. By "observed power" we simply mean the percentage of tests for which the null hypothesis was rejected, calculated over many random trials.

Table 2 compares predicted and observed power for various sample sizes and population effect sizes. The predicted powers in this table are computed exactly, not using the normal approximation. The observed powers are calculated over multiple samples of n topics. Observed power is close to that predicted by the theory, though somewhat higher on average (this may be an artifact of the topic design process or of the particular systems submitted to the track). Note that the standard errors are rather high. The conclusions drawn from a test may vary a lot from sample to sample; a single hypothesis test is therefore not enough to draw strong conclusions.

3.3 Ties in the Sign Test

Ties are trials for which $Y_i = 0$, i.e. there is no measurable difference. Lehmann [7] describes two approaches to ties in the sign test. The usual practice is to discard all trials that resulted in $Y_i = 0$ and reduce n accordingly. In Figure 2(b) we showed how power changes with n; this also shows how

n	h	$1-\beta$	$\widehat{1-\beta}$
25	0.25	0.222	0.246 ± 0.015
	0.35	0.408	0.445 ± 0.018
	0.50	0.727	0.780 ± 0.023
50	0.25	0.478	0.515 ± 0.089
	0.35	0.753	0.816 ± 0.098
	0.50	0.971	0.993 ± 0.012
100	0.25	0.795	0.850 ± 0.083
	0.35	0.971	0.990 ± 0.021
	0.50	1.000	1.000 ± 0.000

Table 2: Predicted and observed rates of detecting significance for varying sample size n and effect size h. Predicted power is denoted $1-\beta$. Observed power is $\widehat{1-\beta}\pm$ one standard error.

power changes as ties become more frequent (as n decreases).

Another approach to ties is to randomly assign them to be successes or failures according to the null hypothesis θ_0. This also decreases the power of the test: the expected number of success after this procedure is $\hat{S} = \sum_{i=1}^{n-n_0} Y_i + \sum_{i=1}^{n_0} \theta_0$ (n_0 the number of ties). If the null hypothesis is true, we are assigning ties to be successes at a rate lower than they would be if they were not ties, and therefore we have less ability to reject the null hypothesis; power decreases.

Of these two methods, the former reduces power less than the latter. In fact, the former method reduces power the least of all possible tie-handling methods [7].

3.4 Uncertainty

By "uncertainty", we mean that there are trials for which we believe that $Y_i = 1$ or $Y_i = 0$, but there is a chance that our measurements are wrong. In IR, uncertainty can come from having incomplete or imperfect relevance judgments. We view uncertainty as being similar to a tie; it is a trial for which there is a measurable difference but the error of that measurement is high. We will denote an uncertain outcome as $\widehat{Y}_i$. Our certainty in Y_i is the probability that $\widehat{Y}_i = 1$ given $Y_i = 1$, i.e. the probability that we have correctly predicted the outcome of the trial. We will call this probability λ.

$$\text{certainty} = \lambda = P(\widehat{Y}_i = 1 | Y_i = 1) \tag{1}$$

Informally, uncertainty = 1− certainty.

3.4.1 Failing to Account for Uncertainty

Failing to account for uncertainty entails using the estimates $\widehat{Y}_i$ and ignoring the uncertainty $1 - \lambda$.

Suppose the null hypothesis is true, i.e. $h = 0$. Given uncertainty $1-\lambda$, what is the probability that we will reject the null hypothesis with significance α and sample size n? The expectation is that there will be equal numbers of successes and failures. Each $\widehat{Y}_i$ we predict to be a success will actually be a failure with probability $1-\lambda$; it will actually be a success with probability λ. Therefore $E[S] = n(\frac{1}{2}\lambda + \frac{1}{2}(1-\lambda)) = \frac{1}{2}n$, which is exactly its expectation when there is no uncertainty at all. Therefore uncertainty does not affect the accuracy of the test.

Now suppose the null hypothesis is false, i.e. $h > 0$ and $\theta > \frac{1}{2}$. If certainty $\lambda > \frac{1}{2}$, the expected number of successes we will observe is $n(\theta\lambda + (1-\theta)(1-\lambda)) \leq n\theta$. Since the expectation is less than what it would be with no uncertainty, power will be reduced.

3.4.2 Accounting for Uncertainty

Above we discussed dealing with ties by randomly assigning them to be positive or negative. This can be generalized to the idea of uncertainty: we can incorporate uncertainty by treating instances that we are uncertain of the true outcome as ties, then assigning them to be success or failures randomly depending on how much uncertainty we have.

For example, suppose the effect size is $h = 0.4$, so $\theta = P(Y_i = 1) = 0.7$ for each trial i. But suppose we have $\widehat{Y}_i$, an estimate of Y_i in which we have certainty $\lambda = P(\widehat{Y}_i = 1 | Y_i = 1) = 0.8$. Then $P(\widehat{Y}_i = 1) = P(\widehat{Y}_i = 1 | Y_i = 1)P(Y_i = 1) + P(\widehat{Y}_i = 1 | Y_i = 0)P(Y_i = 0) = 0.7 \cdot 0.8 + 0.3 \cdot 0.2 = 0.62$ for an effect size of only $h' = 0.24$. Our uncertainty has reduced the effect we can detect from 0.4 to 0.24, thus reducing the power of the test. To make up for the reduction in power, we would need 138 trials rather than 50.

We will define the "adjusted effect size" h' to be the reduced effect size in the presence of uncertainty.

$$h' = \frac{\theta\lambda + (1-\theta)(1-\lambda) - \theta_0}{\theta_0} \tag{2}$$

We can quantify the increase in trials necessary to make up for a loss in power due to uncertainty. Let h be the true effect size ($h > 0$) and h' be the adjusted effect size, with $h' < h$ due to certainty $.5 \leq \lambda < 1$ (if $\lambda < .5$, we may simply flip our prediction and take $\lambda = 1 - \lambda$). Let n be the original sample size. The goal is to find n', the new sample size needed to be able to detect the adjusted effect size with the same power.

Since we want the power to be the same, we need to find the n' that results in the difference in powers being zero:

$$\Phi(Z_\alpha - h\sqrt{n}) - \Phi(Z_\alpha - h'\sqrt{n'}) = 0$$

Since this represents the area under the normal curve from $x_0 = Z_\alpha - h\sqrt{n}$ to $x_1 = Z_\alpha - h'\sqrt{n'}$, it will be minimized when $Z_\alpha - h\sqrt{n} = Z_\alpha - h'\sqrt{n'}$.

Solving for n' gives:

$$n' = n\left(\frac{h}{h'}\right)^2$$

This expression relies on knowing the true effect size, which of course we do not in practice. However, if $\theta_0 = \frac{1}{2}$, as it nearly always would, then

$$\begin{aligned} n' &= n\left(\frac{\theta - \frac{1}{2}}{\theta\lambda + (1-\theta)(1-\lambda) - \frac{1}{2}}\right)^2 \\ &= n\left(\frac{\theta - \frac{1}{2}}{(2\theta\lambda - \lambda) - (\theta - \frac{1}{2})}\right)^2 \\ &= n\left(\frac{\theta - \frac{1}{2}}{2\lambda(\theta - \frac{1}{2}) - (\theta - \frac{1}{2})}\right)^2 \\ n' &= n\left(\frac{1}{2\lambda - 1}\right)^2 \end{aligned} \tag{3}$$

and we have reduced it to an expression that relies only on the original sample size and the uncertainty λ. Though we began with the assumption that we knew the population effect size, the final answer does not depend on that knowledge.

This is the first milestone in this work. As uncertainty increases, the number of trials needed to maintain a given

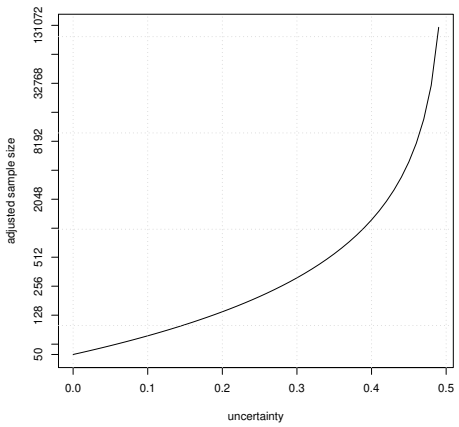


Figure 3: Uncertainty versus sample size n.

level of power increases exponentially. Figure 3 shows how the number of topics must increase to maintain 80% power to detect an effect size of $h = 0.35$ when $n = 50$ and $\alpha = 0.05$.

As we suggested above, a significant source of uncertainty may be incomplete or imperfect relevance judgments. This immediately implies that we can improve the power of the test in two ways: increasing the number of topics or increasing the number of relevance judgments for the extant topics.

4. MEASURING UNCERTAINTY

In our IR example above, each topic is classified as a success or failure depending on the sign of the difference in some evaluation measure. In this section we follow [1] and show how to predict the sign of the difference based on incomplete relevance judgments, and how to estimate the probability that the predicted sign is correct.

We have elected to use the IR evaluation measure average precision (AP). Average precision is a standard evaluation metric that captures both the ability of a system to rank relevant documents highly (precision) and its ability to retrieve relevant documents (recall). It is typically written as the mean precision at the ranks of relevant documents:

$$AP = \frac{1}{|R|} \sum_{i \in R} prec@r(i)$$

where R is the set of relevant documents, and $r(i)$ is the rank of document i. We define ΔAP to be the difference in average precisions between two systems on the same topic. Given an incomplete set of judgments, we can predict ΔAP by assuming anything unjudged is nonrelevant. This is a standard assumption in IR evaluation. However, it gives us no way to assign a probability to our prediction.

Let X_i be a random variable indicating the relevance of document i. If documents are ordered by rank, we can express precision as $prec@i = 1/i \sum_{j=1}^{i} X_j$.

Average precision becomes the quadratic equation

$$AP = \frac{1}{\sum X_i} \sum_{i=1}^{m} X_i / i \sum_{j=1}^{i} X_j$$

where m is the collection size. For a closed form expression of ΔAP, we need to be able to calculate AP when documents are ordered arbitrarily, not necessarily by rank (since the two rankings will most likely be different). To do that, let $a_{ij} = 1/\max\{r(i), r(j)\}$. Then

$$AP = \frac{1}{\sum X_i} \sum_{i=1}^{m} \sum_{j \geq i} a_{ij} X_i X_j$$

To see why this is true, consider a toy example: a list of 3 documents with relevant documents B, C at ranks 1 and 3 and nonrelevant document A at rank 2. Average precision will be $\frac{1}{2}(\frac{1}{1}x_B^2 + \frac{1}{2}x_B x_A + \frac{1}{3}x_B x_C + \frac{1}{2}x_A^2 + \frac{1}{3}x_A x_C + \frac{1}{3}x_C^2) = \frac{1}{2}\left(1 + \frac{2}{3}\right)$ because $x_A = 0, x_B = 1, x_C = 1$. Though the ordering B, A, C is different from the labeling A, B, C, it does not affect the computation.

Doing the same thing for the other list (using b_{ij} rather than a_{ij}), we can then express ΔAP as

$$\Delta AP = \frac{1}{\sum X_i} \sum_{i=1}^{m} \sum_{j \geq i} c_{ij} X_i X_j$$

$$c_{ij} = a_{ij} - b_{ij}$$

We can now see the difference in average precision itself is a random variable with a distribution over all possible assignments of relevance to all documents. This random variable has an expectation, a variance, confidence intervals, and a certain probability of being less than or equal to a given value.

The expectation and variance of ΔAP are:

$$E[\Delta AP] \approx \frac{1}{\sum p_i} \sum \left(c_{ii} p_i + \sum_{j>i} c_{ij} p_i p_j \right)$$

$$Var[\Delta AP] \approx \frac{1}{(\sum p_i)^2} \left(\sum_{i}^{m} c_{ii}^2 p_i (1 - p_i) + \sum_{j>i} c_{ij}^2 p_i p_j (1 - p_i p_j) \right.$$

$$\left. + \sum_{i \neq j} 2 c_{ii} c_{ij} p_i p_j (1 - p_i) + \sum_{k>j \neq i} 2 c_{ij} c_{ik} p_i p_j p_k (1 - p_i) \right)$$

where $p_i = p(X_i = 1)$, the probability that document i is relevant. For simplicity, we set $p_i = \frac{1}{2}$ for all unjudged documents. ΔAP asymptotically converges to a normal distribution with expectation and variance as defined above.[2] This means that we can use the normal cumulative density function to determine the probability that a difference in AP is less than 0.

Assuming topics are independent, we can easily extend this to mean average precision (MAP), the mean of average precisions calculated for a set of topics T. MAP is also normally distributed with expectation and variance:

$$\mathcal{E}MAP = \frac{1}{T} \sum_{t \in T} E[AP_t] \tag{4}$$

$$\mathcal{V}MAP = \frac{1}{T^2} \sum_{t \in T} Var[AP_t]$$

And we define $\Delta MAP = MAP_1 - MAP_2$ analogously to ΔAP. ΔMAP has an expectation and variance as well.

We will define confidence to be

$$\text{confidence} = P(\Delta MAP < 0) = \Phi \left(\frac{-E[\Delta MAP]}{\sqrt{Var[\Delta MAP]}} \right)$$

[2]These are actually approximations to the true expectation and variance, but the error is a negligible $\mathcal{O}(m2^{-m})$.

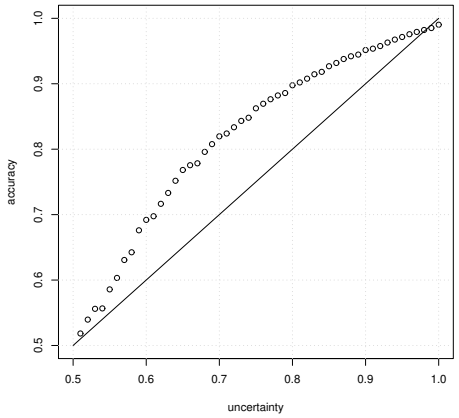


Figure 4: Uncertainty in $\widehat{\Delta MAP}$ versus actual ΔMAP.

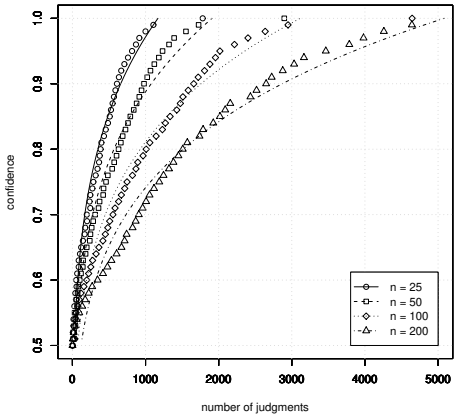


Figure 5: Number of judgments required to reach uncertainty levels for varying sample sizes. The fit lines are shown as well.

4.1 Certainty and Confidence

We defined certainty above as $P(\widehat{Y}_i|Y_i)$. We would like to connect the idea of confidence to the idea of certainty. There is a rather natural connection: let $\widehat{Y}_i = sgn(E[\Delta AP_i])$. We will define $\lambda = P(\widehat{Y}_i = 1|Y_i = 1) = P(\Delta MAP > 0)$. The reason for using ΔMAP to assign the probability rather than ΔAP is that topics are assumed to have been drawn i.i.d. from a population in which we have λ certainty on every member. Certainties may vary from topic to topic, but the topic certainties are samples from a distribution with expectation equal to the population uncertainty.

In order to use confidence as certainty, we would like to see that if $sgn(\Delta AP) = 1$ and certainty is λ, then the proportion of pairs for which $sgn(E[\Delta AP]) = 1$ is at least λ. To test this, we used the Robust runs from Section 3.2.1.

For each pair of runs over the full set of 249 topics, we judge a "pool" of depth k (the top-ranked k documents by both runs for all topics), from $k = 1$ to 100. After each increment of k, we estimate the difference in average precision $\widehat{Y}_i = sgn(E[\Delta AP])$ and the confidence $P(\Delta MAP > 0)$ (with probability of relevance $p_i = \frac{1}{2}$).

The results are shown in Figure 4. The solid line is what we would see if confidence exactly predicted accuracy; since our points are uniformly above that line, it seems that confidence meets our requirements for a measure of uncertainty. Therefore we use "confidence" and "certainty" interchangeably for the remainder of this work.

4.2 Judgments and Confidence

The evaluation in the previous section gives us data to estimate the number of judgments it takes to reach increasing confidence levels with increasing numbers of topics.

Figure 5 shows the average number of judgments needed to achieve increasing confidence levels. Confidence levels may fluctuate, so that after achieving 70% confidence, a few more judgments cause confidence to drop below 70%. The judgments are the average minimum number required for confidence, i.e. the number of judgments made when confidence level λ was first achieved. This models an assessor that stops judging the first time confidence reaches a given threshold.

The figure shows an exponential relationship between judgments and confidence. It also shows a relationship between the number of topics and the number of judgments.

We will fit a curve to these plots to estimate the relationship. Define the estimated number of judgments needed to reach λ confidence for n topics as

$$\begin{aligned}\hat{j}(\lambda, n) &= e^{\gamma_0 + \gamma_1 \log \lambda + \gamma_2 \log n} \\ &= e^{\gamma_0} \lambda^{\gamma_1} n^{\gamma_2} \end{aligned} \quad (5)$$

We can fit this model using regression. Since the number of judgments is a count, we do not want to use an ordinary least squares estimation, which could lead to predictions that are less than one. Instead, we will fit a generalized linear model with a Poisson link function. This guarantees that all predictions will be at least 1. For more information on generalized linear models and Poisson regression, we refer the reader to Faraway [4] or Venables & Ripley [11].

The result of fitting the Poisson regression to our data is is

$$\hat{j}(\lambda, n) = e^{4.79} \lambda^{5.43} n^{0.71} \quad (6)$$

The R^2 for this model is 0.95, so it is a good fit.

The fact that it takes exponentially many judgments to increase confidence suggests that it may be more cost-effective to obtain a large number of topics with a few judgments for each, rather than judging a large number of judgments for a small number of topics. But recall from Figure 3 that the number of topics needed also increases exponentially as uncertainty increases. Therefore we need a cost-benefit analysis to tell us what to do.

5. USING UNCERTAINTY TO EXPERIMENT

Given our equation for the new sample size needed to maintain power in the presence of uncertainty (Eq. 3) and our model for estimating the number of judgments (Eq. 5), we can figure out the most cost-beneficial confidence level to aim for.

We will define a cost function C associated with a level of confidence λ and original sample size n. The cost is the

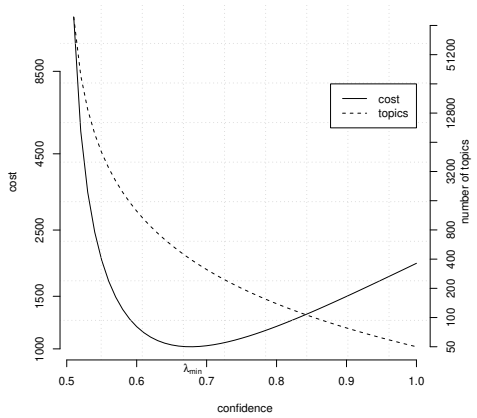


Figure 6: Certainty/confidence versus estimated cost, and certainty versus adjusted sample size.

total cost of developing n' topics plus the total cost of acquiring the predicted number of relevance judgments needed to reach confidence λ. Let C_t be the cost of developing a topic. Let C_j be the cost of judging one document. Suppose our sample size n has been selected in advance (or selected for us).

$$\begin{aligned} C(\lambda, n) &= C_t n' + C_j \hat{j}(\lambda, n') \\ &= C_t n' + C_j e^{\gamma_0} \lambda^{\gamma_1} n'^{\gamma_2} \\ &= C_t n \left(\frac{1}{2\lambda - 1}\right)^2 + C_j e^{\gamma_0} \lambda^{\gamma_1} \left(n \left(\frac{1}{2\lambda - 1}\right)^2\right)^{\gamma_2} \end{aligned} \tag{7}$$

which is obtained by substituting Eq. 3 for n' and Eq. 5 for $\hat{j}(\lambda, n')$. Figure 5 shows how cost changes with target level of confidence λ (we have set $C_j = 1$ and $C_t = 0$ for this example).

We wish to minimize Eq. 7 with respect to λ. There is no analytic minimum, but an approximate minimum can be found easily by binary search over discrete values of λ.

Estimating the cost of developing a topic is difficult. The fact that topics can be reused more easily than relevance judgments (many TREC topics are "portable" over collections) should be considered, as should the fact that the same topics can be used to evaluate wildly diverse retrieval systems that may retrieve complete different documents and therefore need completely different judgments. Sometimes topics are developed by a third party and given to us, or sampled from a query log at very little cost.

When $C_t = 0$, the cost function has an analytic minimum. We differentiate with respect to lambda.

$$\frac{dC}{d\lambda} = C_j \left(\frac{\gamma_1}{\lambda} - \frac{4\gamma_2}{2\lambda - 1}\right) e^{\gamma_0} \lambda^{\gamma_1} \left(n \left(\frac{1}{2\lambda - 1}\right)^2\right)^{\gamma_2}$$

Equating this to zero and solving for λ gives

$$\lambda = \frac{\gamma_1}{2\gamma_1 - 4\gamma_2} \tag{8}$$

which means the level of confidence that minimizes cost is independent of the cost of making judgments, and independent of the original sample size.

Instead of using n in our cost function, we could include Type I and Type II error rates and associated costs C_α and C_β. We would then minimize along several dimensions (λ, α, β). Since estimating the cost of false positives and false negatives is tricky and to some degree personal, we do not explore this further.

5.1 Example Usage

Suppose we wish to be able to detect an effect size of 0.5 with 80% power. As Table 2 shows, about $n = 25$ topics is an appropriate sample size if there is no uncertainty due to relevance judgments, i.e. we have a full set of judgments and little to no assessor disagreement.

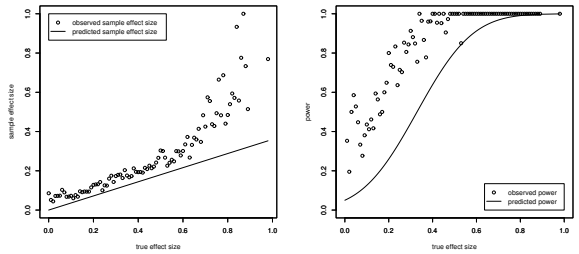
If we want no uncertainty, we must have $\lambda = 1$. Plugging into our cost function (7) (assuming topics are free) gives $C(1, 25) \approx 1180$ relevance judgments. Cost can be greatly reduced, though; if confidence is reduced to 80%, cost is reduced to $C(0.8, 25) \approx 914$. Plugging the coefficients from the model we trained above into Eq. 8, we find that the minimum cost is achieved when confidence is 68%: $C(0.68, 25) \approx 620$. The adjusted sample size needed to maintain the power of the test is $n' = 25 \left(\frac{1}{2 \cdot .68 - 1}\right)^2 \approx 192$, so we estimate that judging 192 topics to 68% confidence costs nearly half as much as judging 25 topics to 100% confidence, without reducing the power of the test at all. Judging a pool of depth 100 for 25 topics would require 4,161 judgments on average; our cost is 85% less than that.

To test whether our predictions matched reality, we picked 2,000 pairs of Robust systems at random. For each pair, we evaluated 192 topics to 68% confidence. We also evaluated 25 topics to 100% confidence and a pool of depth 100 for 25 topics.

The actual number of judgments needed to reach 68% confidence ranged from a minimum of 44 to a maximum of 10,164, with a mean of 794 but a median of only 357. About 70% of the trials required fewer than our prediction of 620 judgments. There is a great deal of variance in the number of judgments needed, but indeed it required 1,000 fewer judgments on average for 192 topics than 25. It required 3,406 fewer judgments on average to judge 192 topics than to judge a pool of depth 100 for 25 topics.

What about the power of the adjusted sample size? We should be able to detect an effect size of 0.5 about 80% of the time. In fact, 83% of the pairs with a "true" effect size of 0.5 were found to be significant with $n' = 192$ topics.

We calculated "observed adjusted effect size" by counting the number of topics for which $\widehat{Y}_i$ was positive and the number for which $\widehat{Y}_i$ was nonzero. These observations are compared to the predicted adjusted effect size (using Eq. 2) in Figure 7(a). We generally underestimate the adjusted effect size; this is most likely because confidence underpredicts accuracy (Figure 4). We also calculated "observed power" by counting the number of trials for which the null hypothesis is rejected at $\alpha = 0.05$. This is shown in Figure 7(b), along with the predicted power of 25 topics and 100% confidence in each. Empirically, using 192 topics with 68% confidence actually has *more* power than using 25 topics with 100% confidence; again, this is most likely due to confidence underpredicting accuracy. Recall that since our "population" only consists of 249 topics, there is some effect between every pair, so the null hypothesis is always false in this data;



(a) Observed effect size. (b) Observed power.

Figure 7: Effect size and power for 192 topics at 68% confidence. The solid lines are predicted by our analysis; the points are empirical performance.

though the points at the left end of the plot are high, they are not Type I errors. In reality, there would be cases in which the null hypothesis is true.

5.2 Uncertainty in Experimental Design

In this section we show how to use uncertainty in experimental design. We assume that we have a new retrieval task. There are no existing topics, no existing relevance judgments. We would like to perform an experiment, but have limited resources. We need to find the minimum-cost parameters that will not cost us any of our desired power.

We do not have access to our previously-trained prediction model $\hat{j}$, so the first thing we need to do is train one. To do that, we first run a pilot study. Pilot studies are common in disciplines in which the cost of running tests is high; Jensen [6] also proposed the use of pilot tests to determine the effect of uncertain judgments.

We develop 10 topics for the pilot study. The topics are submitted to the two runs we wish to compare. All 10 topics are judged until there is 100% confidence in ΔMAP. We keep track of the minimum number of judgments required to reach each intermediate confidence level. This does not yet give us data to train $\hat{j}$; for that we need judgment counts and confidence levels for varying numbers of topics.

We can get training data for $n < 10$ by simulating an evaluation on a subset of the topics using the judgments we have just made. We can get training data for $n > 10$ by sampling with replacement from our set of 10 topics, then simulating evaluation over the larger set. The results of these simulations are used to train $\hat{j}$.

Now, using n and $\hat{j}$, we may find the minimum-cost number of topics and judgments.

The pilot study serves two purposes: one, to see if the experiment is worth continuing; if it is, then to estimate the amount of work necessary to carry it to completion.

5.2.1 Experiment

To experiment with this pilot study, we will again use Robust systems and topics. Of course, they do not represent a new task or new topics, but they provide a useful "truth" to compare against.

The ten topics randomly selected for the pilot study were 315, 350, 367, 393, 442, 602, 610, 670, 675, and 681. As described above, a pair of systems (selected randomly) is first evaluated over these 10 topics. The model $\hat{j}$ is trained and the minimum-cost λ and n' found using Eq. 7. We then randomly select n' new topics (excluding the ones from the pilot study) and evaluate to confidence λ on those.

We do two experiments. For the first, the first 249 topics are free ($C_t = 0$) but any more than that cost $20C_j$; this would be like receiving 249 topics from NIST but no relevance judgments. For the second, all topics cost 20 times as much as relevance judgments. If the optimal number of topics is greater than 249, we create new topics by sampling with replacement from the existing topics. They will still be treated as different topics, so if we have duplicated topic 301, judgments for 301 will not count towards 301′. (As it turns out, we never had to do this.)

As an example, consider systems `polyudp5` and `NLPR04SemLM`. For the first 10 topics it requires 1,587 relevance judgments to reach 100% confidence in the difference between them. We simulate evaluating one topic, two topics, and so on; the resulting model is $\hat{j}(\lambda, n) = \exp(5.02 + 5.6 \log \lambda + 0.91 \log n)$. Using this model, we predict the minimum cost to be achieved with 51 topics and 85% confidence (1625 relevance judgments). It ends up taking 836 judgments to reach 85% confidence on 51 topics, so we overestimated the cost. We have money left over for a pizza party for our assessors!

5.2.2 Results

For the first experiment, the extra cost of using more than 249 topics resulted in the model never selecting more than 216. Cost, therefore, is equivalent to the number of judgments. On average, over 2,000 trials, cost was predicted to be 1,492 judgments to reach 75% confidence over 157 topics. In actuality, it required 950 judgments on average to reach the target confidence. For comparison, the number of judgments needed to reach 100% confidence over 25 topics was 1,781; this is a 47% decrease in the number of judgments.

Since 542 fewer judgments were required than predicted, over many experiments our cost function will tend to perform better than expected. However, more than half the trials required more judgments than predicted. The correlation between the predicted and actual number of judgments is 0.43, indicating that the model is doing a reasonable job but could be better. This is partially affected by the pilot sample. If the pilot sample is "harder" than the population, we may consistently overestimate the number of judgments required; if the pilot sample is "easier", we may consistently underestimate the number of judgments required. In this case it seems our pilot sample was a bit easier than the population. Using a larger pilot sample could produce more accurate predictions, but of course would require a greater start-up cost in the pilot study.

The start-up cost is the number of judgments needed to reach 100% confidence on the pilot sample of 10 topics. Over 2,000 trials, the mean start-up cost was 871 judgments. No trial required more than 2,000 judgments, which again points to our pilot sample being easier than the population. About 63% of trials had a start-up cost of 1,000 judgments or fewer.

Figure 8 shows the observed power as well as the power for 25 topics at 100% confidence. Power is high. In fact, we have outperformed our predictions, making up for the trials in which we underpredicted the number of judgments.

For the second experiment, in which topics cost 20 times as much as relevance judgments, it turns out that it is often most cost-beneficial to judge 25 topics to 100% confidence. In over half the trials, it was more cost-effective to judge

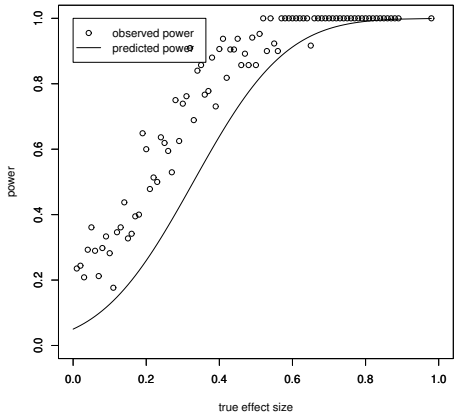


Figure 8: Observed power over 2,000 trials training a prediction model $\hat{j}$ using data from a pilot study. The solid line is the power of using 25 topics at 100% confidence; the points are the empirical result of using more topics with less confidence (fewer judgments).

25 topics to 100% confidence. It was never cost-effective to judge more than 74 topics. Of course, these topics could then be reused, and would be free for the next experiment.

Since fewer topics were used, the number of judgments was higher than the previous experiment in which topics were free. On average, 1438 judgments were made to reach 0.97 confidence over 29.2 topics. Including the cost of developing topics, total mean actual cost was 2022. This was greater than the predicted cost of 1858, though this time fewer than half of the trials required more judgments than predicted.

The observed power is actually a little worse than predicted (not shown). This is because smaller sets of topics are more prone to errors due to sampling, even with 100% confidence. This is further reinforcement that more topics is superior.

6. CONCLUSIONS AND FUTURE WORK

We have proved that a large number of topics with a few relevance judgments for each is as good for evaluation purposes as a small number of topics with a lot of relevance judgments for each. Furthermore, we have shown that the former is much less expensive than the latter: 50% less assessor effort compared to having 100% confidence in each topic; 80% less assessor effort compared to judging a pool of depth 100 for each topic.

The biggest weakness of our cost function is the model for predicting the number of judgments needed. There is such a huge amount of variance over systems and topics that it is very difficult to predict with good accuracy. We have some ideas for improving the accuracy, however: preliminary experiments suggest that measuring the similarity between ranked lists and including it as a feature in the model improves predictions substantially. Additionally, rather than train using the average minimum number of judgments required to get to a given confidence level, we could train an "upper bound" model using quantiles of judgments. Preliminary experiments with quantile regression models are encouraging.

All of our experiments were done using pairs of systems. In reality, researchers would often have multiple systems to evaluate. Hypothesis testing in these situations becomes more difficult, as errors become more frequent simply by chance. This is known as the "multiple testing problem". It requires ad hoc adjustments to Type I and Type II error rates, and is beyond the scope of this work.

Finally, an obvious direction for future work is to analyze the Wilcoxon sign rank test and the t-test in the same way. These tests are perhaps more widely used in IR experimentation than the sign test, and generally have more power to detect the types of differences we are interested in [9].

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR001-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] B. Carterette, J. Allan, and R. K. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 268–275, 2006.

[2] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Earlbaum Associates, 1988.

[3] G. V. Cormack and T. R. Lyman. Power and bias of subset pooling strategies. In *Proceedings of SIGIR*, pages 837–838, 2007.

[4] J. Faraway. *Extending the Linear Model with R*. Chapman & Hall, 2005.

[5] J. M. Hoenig and D. M. Heisey. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):11–6, 2001.

[6] E. C. Jensen. *Repeatable Evaluation of Information Retrieval Effectiveness in Dynamic Environments*. PhD thesis, Illinois Institute of Technology, 2006.

[7] E. L. Lehmann. *Testing Statistical Hypotheses*. Springer, 1997.

[8] M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of SIGIR*, pages 186–193, 2005.

[9] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM*, 2007. To appear.

[10] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

[11] W. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 2003.

[12] E. M. Voorhees. Overview of the 2004 trec robust track. In *13th TREC*, 2004.

[13] D. Wackerly, W. Mendenhall, and R. L. Sheaffer. *Mathematical Statistics With Applications*. P W S Publishers, 5th edition, 1995.

[14] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.