A Comparison of Statistical Significance Tests for Information Retrieval Evaluation

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

ABSTRACT

Information retrieval (IR) researchers commonly use three tests of statistical significance: the Student's paired t-test, the Wilcoxon signed rank test, and the sign test. Other researchers have previously proposed using both the bootstrap and Fisher's randomization (permutation) test as nonparametric significance tests for IR but these tests have seen little use. For each of these five tests, we took the ad-hoc retrieval runs submitted to TRECs 3 and 5-8, and for each pair of runs, we measured the statistical significance of the difference in their mean average precision. We discovered that there is little practical difference between the randomization, bootstrap, and t tests. Both the Wilcoxon and sign test have a poor ability to detect significance and have the potential to lead to false detections of significance. The Wilcoxon and sign tests are simplified variants of the randomization test and their use should be discontinued for measuring the significance of a difference between means.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

Statistical significance, hypothesis test, sign, Wilcoxon, Student's t-test, randomization, permutation, bootstrap

1. INTRODUCTION

A chief goal of the information retrieval (IR) researcher is to make progress by finding better retrieval methods and avoid the promotion of worse methods. Given two information retrieval (IR) systems, how can we determine which one is better than the other? A common batch-style experiment is to select a collection of documents, write a set of topics, and create relevance judgments for each topic and then measure the effectiveness of each system using a metric like the mean average precision (MAP). TREC typifies this style of evaluation [20].

We know that there is inherent noise in an evaluation. Some topics are harder than others. The assessors hired to judge relevance of documents are human and thus open to variability in their behavior. And finally, the choice of document collection can affect our measures.

We want to promote retrieval methods that truly are better rather than methods that by chance performed better given the set of topics, judgments, and documents used in the evaluation. Statistical significance tests play an important role in helping the researcher achieve this goal. A powerful test allows the researcher to detect significant improvements even when the improvements are small. An accurate test only reports significance when it exists.

An important question then is: what statistical significance test should IR researchers use?

We take a pragmatic approach to answering this question. If two significance tests tend to produce nearly equivalent significance levels (p-values), then to the researcher there is little practical difference between the tests. While the underlying fundamentals of the tests may be very different, if they report the same significance level, the fundamental differences cease to be practical differences.

Using the runs submitted to five TREC ad-hoc retrieval evaluations, we computed the significance values for the Student's paired t, Wilcoxon signed rank, sign, shifted bootstrap, and randomization tests. Comparing these significance values we found that:

- Student's t, bootstrap, and randomization tests largely agree with each other. Researchers using any of these three tests are likely to draw the same conclusions regarding statistical significance of their results.
- The Wilcoxon and sign tests disagree with the other tests and each other. For a host of reasons that we explain, the Wilcoxon and sign tests should no longer be used by IR researchers.

We also came to the following conclusions as part of our study:

- A test should test the same statistic that a researcher reports. Thus, the t-test is only appropriate for testing

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

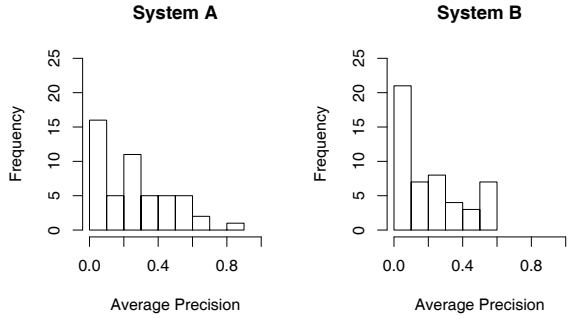

Figure 1: The distribution of 50 average precision scores for two example IR systems submitted to TREC 3. The mean average precision (MAP) of system A is 0.258 and the MAP of system B is 0.206.

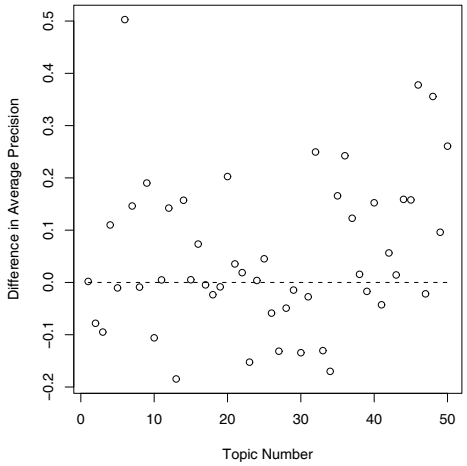

Figure 2: The per topic differences in average precision of the same two systems as in Figure 1.

the difference between means. Both the randomization and bootstrap can use any statistic.

- Based on the tests' various fundamentals, we recommend the randomization test as the preferred test in all cases for which it is applicable.

Other researchers have studied the use of significance tests as part of IR evaluation [5, 6, 10, 16, 17, 18, 21], but we know of no other work that looks at all of these tests or takes our pragmatic, comparative approach.

2. SIGNIFICANCE TESTING

As Box, Hunter, and Hunter [1] explain, a significance test consists of the following essential ingredients:

1. A test statistic or criterion by which to judge the two systems. IR researchers commonly use the difference in mean average precision (MAP) or the difference in the mean of another IR metric.

2. A distribution of the test statistic given our *null hypothesis*. A typical null hypothesis is that there is no difference in our two systems.

3. A significance level that is computed by taking the value of the test statistic for our experimental systems and determining how likely a value that large or larger could have occurred under the null hypothesis. This probability of the experimental criterion score given the distribution created by null hypothesis is also known as the *p-value*.

When the significance level is low, the researcher can feel comfortable in *rejecting the null hypothesis*. If the null hypothesis cannot be rejected, then the difference between the two systems may be the result of the inherent noise in the evaluation.

To make our discussion more concrete, we will use two actual runs submitted to TREC 3 as an example. On the 50 topics of TREC 3, system A had a MAP of 0.258 and system B had a MAP of 0.206. Figure 1 shows the distribution of average precision scores for systems A and B.

We know that a large amount of the variability in the scores on an IR evaluation comes from the topics. Each system produces a score for each topic and on a per-topic basis we obtain *matched pairs* of scores. All of the tests we consider evaluate significance in light of this paired design, which is common to most batch-style IR experiments. Figure 2 shows the per topic differences between the two example systems.

As measured by mean average precision, system A performed 20.1% better than B, but is this a statistically significant improvement? We have already chosen our criterion by which to judge the difference of the two systems – difference in mean average precision. We next need to form a *null hypothesis* and determine whether we can reject the null hypothesis.

Each of the following significance tests has its own criterion and null hypothesis. The randomization and bootstrap tests can use whatever criterion we specify while the other tests are fixed in their test statistic. While there are fundamental differences in the null hypotheses, all of the tests aim to measure the probability that the experimental results would have occurred by chance if systems A and B were actually the same system.

2.1 Randomization Test

For Fisher's randomization test [1, 4, 8, 9], our null hypothesis is that system A and system B are identical and thus system A has no effect compared to system B on the mean average precision for the given topics, corpora, and relevance judgments.

If system A and system B are identical, we can imagine that there is some system N that produced the results for A and B. For example, on one topic, system A had an average precision (AP) of 0.499 and system B had an AP of 0.577. Under the null hypothesis, system N produced both results and we merely labeled one as being produced by system A and the other by system B. To generate the results for all 50 topics, we asked system N to produce two results for each topic and we labeled one of them as produced by A and the other by B.

Thus, if system A and system B are identical, then we can think of them as merely labels applied to the scores produced by system N. The decision to label one score for a topic as

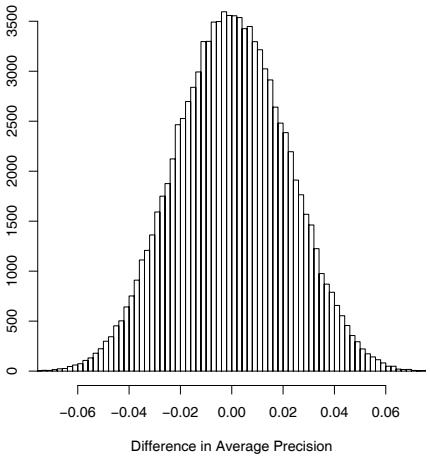

Figure 3: The distribution of 100,000 differences in mean average precision between random permutations of system A and B.

produced by system A or B is arbitrary. In fact, since there are 50 topics, there are 2^{50} ways to label the results under the null hypothesis. One of these labelings is exactly the labeling of the example that produced MAPs of 0.258 for system A and 0.206 for system B.

Under the null hypothesis, any permutation of the labels is an equally likely output. We can measure the difference between A and B for each permutation. If we created each of the 2^{50} permutations, we could measure the number of times a difference in MAP was as great or greater than the difference we measured in our example ($0.258 - 0.206 = 0.052$). This number divided by 2^{50} would be the exact one-sided *p-value* or *achieved significance level* for the null hypothesis. Doing such a test of the null hypothesis is known as a *randomization test* or *permutation test*. If we measured the number of times the absolute value of the difference was as great or greater than the measured difference, we would have the exact two-sided p-value.

Computing 2^{50} permutations takes even a fast computer longer than any IR researcher is willing to wait. An alternative is to sample and produce a limited number of random permutations. The more samples, the more accurate will our estimate of the p-value be. We will discuss the number of samples needed in Section 3.

Returning to our example, we created 100,000 random permutations of system A and B and measured the difference in mean average precision for each arrangement. Figure 3 shows the distribution of differences. Our example's MAP difference is 0.052. Of the 100,000 measured differences, 689 are ≤ -0.052 and 691 are ≥ 0.052. This gives us a two-sided p-value of $(689+691)/100000 = 0.0138$. This shows that the difference of 0.052 is unlikely and thus we should reject the null hypothesis and report that system A has achieved a statistically significant improvement over system B.

Before the era of cheap computer power, the randomization test was impractical for all but the smallest experiments. As such, statisticians created significance tests that replaced the actual score differences with the ranks of the scores [2]. Of these tests, IR researchers have most widely used the Wilcoxon signed rank test.

2.2 Wilcoxon Signed Rank Test

The null hypothesis of the Wilcoxon signed rank is the same as the randomization test, i.e. systems A and B have the same distribution [13].

Whereas the randomization test can use any test statistic, the Wilcoxon test uses a specific test statistic. The Wilcoxon test statistic takes the paired score differences and ranks them in ascending order by absolute value. The sign of each difference is given to its rank as a label so that we will typically have a mix of "negative" and "positive" ranks. For a two-sided test, the minimum of the sums of the two sets of ranks is the test statistic. Differences of zero and tied differences require special handling [13].

The Wilcoxon test statistic throws away the true differences and replaces them with ranks that crudely approximate the magnitudes of the differences. This loss of information gained computational ease and allowed the tabulation of an analytical solution to the distribution of possible rank sums. One refers the test statistic to this table to determine the p-value of the Wilcoxon test statistic. For sample sizes greater than 25, a normal approximation to this distribution exists [13].

For our example, the Wilcoxon p-value is 0.0560. This is significantly larger than the randomization test's 0.0138. While we would likely judge the systems to be significantly different given the randomization test, we would likely come to the opposite conclusion using the Wilcoxon test.

Of note is that the null hypothesis distribution of the Wilcoxon test statistic is the same distribution as if this test statistic was used for the randomization test [11]. Thus we can see the dramatic affect that choosing a different test statistic can have for a statistical test.

Wilcoxon's test made sense when Wilcoxon presented it in 1945 as a test to "obtain a rapid approximate idea of the significance of the differences" [22]. Given that IR researchers will use a computer to compute their significance tests, there seem to be only disadvantages to the test compared to the randomization test; the randomization test can use the actual test statistic of concern such as the difference in mean.

Wilcoxon believed that one of the advantages of his test was that its utilization of magnitude information would make it better than the sign test, which only retains the direction of the difference [22].

2.3 Sign Test

Like the randomization and Wilcoxon tests, the sign test has a null hypothesis that systems A and B have the same distribution [13].

The test statistic for the sign test is the number of pairs for which system A is better than system B. Under the null hypothesis, the test statistic has the binomial distribution with the number of trials being the total number of pairs. The number of trials is reduced for each tied pair.

Given that IR researchers compute IR metrics to the precision available on their computers, van Rijsbergen proposed that a tie should be determined based on some set absolute difference between two scores [19]. We will refer to this variant of the test as the sign minimum difference test and abbreviate it as the *sign d.* test in tables and figures. We used a minimum absolute difference of 0.01 in average precision for our experiments with the sign minimum difference test.

For our example, system A has a higher average precision than system B on 29 topics. The two-sided sign test with 29 successes out of 50 trials has a p-value of 0.3222. The sign minimum difference test with the minimum difference set to 0.01 has 25 successes out of 43 trials (seven "ties") and a p-value of 0.3604. Both of these p-values are much larger than either the p-values for the randomization (0.0138) or Wilcoxon (0.0560) tests. An IR researcher using the sign test would definitely fail to reject the null hypothesis and would conclude that the difference between systems A and B was a chance occurrence.

While the choice of 0.01 for a minimum difference in average precision made sense to us, the sign minimum difference test is clearly sensitive to the choice of the minimum difference. If we increase the minimum difference to 0.05, the p-value drops to 0.0987.

The sign test's test statistic is one that few IR researchers will report, but if an IR researcher does want to report the number of successes, the sign test appears to be a good candidate for testing the statistical significance.

The sign test, as with the Wilcoxon, is simply the randomization test with a specific test statistic [11]. This can be seen by realizing that the null distribution of successes (the binomial distribution) is obtained by counting the number of successes for the 2^N permutations of the scores for N trials, where for our IR experiments N is the number of topics.

As with the Wilcoxon, given the modern use of computers to compute statistical significance, there seem to only be disadvantages to the use of the sign test compared to the randomization test used with the same test statistic as we are using to measure the difference between two systems.

2.4 Bootstrap Test – Shift Method

As with the randomization test, the bootstrap shift method significance test is a distribution-free test. The bootstrap's null hypothesis is that the scores of systems A and B are random samples from the same distribution [4, 8, 14]. This is different than the randomization test's null hypothesis that makes no assumptions about random sampling from a population.

The bootstrap tries to recreate the population distribution by sampling with replacement from the sample. For the shift method, we draw pairs of scores (topics) with replacement from the scores of systems A and B until we have drawn the same number of pairs as in the experiment. For our example the number of topics is 50. Once we have our 50 random pairs, we compute the test statistic over this new set of pairs, which for our example is the difference in the mean average precision. The bootstrap can be used with any test statistic.

We repeat this process B times to create the bootstrap distribution of the test statistic. As we did with the randomization test, we set $B = 100{,}000$, which should be more than adequate to obtain an accurate bootstrap distribution.

The bootstrap distribution is not the same as the null hypothesis distribution. The shift method approximates the null hypothesis distribution by assuming the bootstrap distribution has the same shape as the null distribution. The other tests we examine do not make any similar guesses. The other tests directly determine the null hypothesis distribution.

We then take the bootstrap distribution and shift it so that its mean is zero. Finally, to obtain the two-sided p-value, we determine the fraction of samples in the shifted distribution that have an absolute value as large or larger than our experiment's difference. This fraction is the p-value.

For our example of system A compared to system B, the bootstrap p-value is 0.0107. This is comparable to the randomization test's p-value of 0.0138.

2.5 Student's Paired t-test

In some ways the bootstrap bridges the divide between the randomization test and Student's t-test. The randomization test is distribution-free and is free of a random sampling assumption. The bootstrap is distribution-free but assumes random sampling from a population. The t-test's null hypothesis is that systems A and B are random samples from the same normal distribution [9].

The details of the paired t-test can be found in most statistics texts [1]. The two-sided p-value of the t-test is 0.0153, which is in agreement with both the randomization (0.0138) and bootstrap (0.0107) tests.

In 1935 Fisher [9] presented the randomization test as an "independent check on the more expeditious methods in common use." The more expeditious methods that he refers to are the methods of Student's t-test. He was responding to criticisms of the t-test's use of the normal distribution in its null hypothesis. The randomization test provided a means to test "the wider hypothesis in which no normality distribution is implied." His contention was that if the p-value produced by the t-test was close to the p-value produced by the randomization test, then the t-test could be trusted. In practice, the t-test has been found to be a good approximation to the randomization test [1].

2.6 Summary

For our example, the randomization test has a p-value of 0.0138. The Wilcoxon signed rank test's p-value is 0.0560. The sign test has a p-value of 0.3222 and the sign minimum difference test has a p-value of 0.3604. The bootstrap has a p-value of 0.0107, and the paired t-test's p-value is 0.0153. All p-values are for two-sided tests.

If a researcher decides to declare p-values less than 0.05 as significant, then only the randomization, bootstrap, and t tests were able to detect significance. The Wilcoxon and sign p-values would cause the researcher to decide that system A did not produce a statistically significant increase in performance over system B.

If the Wilcoxon and sign test tend to produce poor estimates of the significance of the difference between means, a researcher using the Wilcoxon or sign test is likely spend a lot longer searching for methods that improve retrieval performance compared to a researcher using the randomization, bootstrap, or t test.

We next describe our experiments to measure the degree to which the various tests agree with each other.

3. METHODS AND MATERIALS

In this section, we describe the details of the data used and important specifics regarding our experiment.

We took the ad-hoc retrieval runs submitted to TRECs 3 and 5–8 and for each pair of runs, we measured the statistical significance of the difference in their mean average precision. This totaled 18820 pairs with 780, 1830, 2701, 5253, and 8256 pairs coming from TRECs 3, 5-8 respectively.

We computed all runs' scores using trec_eval [3]. We specified the option -M1000 to limit the maximum number of documents per query to 1000.

We measured statistical significance using the Student's paired t-test, Wilcoxon signed rank test, the sign test, the sign minimum difference test, the bootstrap shift method, and the randomization test. The minimum difference in average precision was 0.01 for the sign minimum difference test. All reported p-values are for two-sided tests.

We used the implementations of the t-test and Wilcoxon in R [15] (`t.test` and `wilcox.test`). We implemented the sign test in R using R's binomial test (`binom.test`) with ties reducing the number of trials.

We implemented the randomization and bootstrap tests ourselves in C++. Our program can input the relational output of trec_eval.

Since we cannot feasibly compute the 2^{50} permutations required for an exact randomization test of a pair of TREC runs, each scored on 50 topics, we randomly sampled from the permutations. The coefficient of variation of the estimated p-value, $\hat{p}$ as shown by Efron and Tibshirani [8] is:

$$cv_B(\hat{p}) = \left(\frac{(1-p)/p}{B}\right)^{1/2}$$

where B is the number of samples and p is the actual one-sided p-value. The coefficient of variation is the standard error of the estimated p-value divided by the mean. For example, to estimate a p-value of 0.05 with an error of 10% requires setting $p = 0.05$ and $B = 1901$ to produce a $cv_B(\hat{p}) = 0.1$. To estimate the number of samples for a two sided test, we divide p in half.

For our comparative experiments, we used 100,000 samples. For a set of experiments in the discussion, we used 20 million samples to obtain a highly accurate p-value for the randomization test. The p-value for the randomization test with 100K samples differs little from the value from 20M samples.

With 100K samples, a two-sided 0.05 p-value is computed with an estimated error of 2% or ±0.001 and a 0.01 p-value has an error of 4.5% or ±0.00045. This level of accuracy is very good.

With $B = 20 \times 10^6$, an estimated two-sided p-value of 0.001 should be accurate to within 1% of its value. As the estimated p-value get larger, they become more accurate estimates. For example, a 0.1 p-value will be estimated to within 0.01% or ±0.0001 of its value. Thus, even with the small p-values that concern most researchers, we will have calculated them to an estimated accuracy that allows us to use them as a gold standard to judge other tests with the same null hypothesis.

On a modern microprocessor, for a pair of runs each with 50 topics, our program computes over 511,000 randomization test samples per second. Thus, we can compute a randomization test p-value for a pair of runs in 0.2 seconds using only 100K samples.

We do not know how to estimate the accuracy of the bootstrap test's p-values given a number of samples, but 100K samples is 10 to 100 times more samples than most texts recommend. Wilbur [21] and Sakai [16] both used 1000 samples for their bootstrap experiments.

Selection of a random number generator (RNG) is important when producing large numbers of samples. A poor RNG will have a small period and begin returning the same sequence of random numbers. We used Matsumoto and Nishimura's Mersenne Twister RNG [12], which is well suited for Monte Carlo simulations given its period of $2^{19937} - 1$.

	rand.	t-test	boot.	Wilcx.	sign	sign d.
rand.	-	0.007	0.011	0.153	0.256	0.240
t-test	0.007	-	0.007	0.153	0.255	0.240
boot.	0.011	0.007	-	0.153	0.258	0.243
Wilcx.	0.153	0.153	0.153	-	0.191	0.165
sign	0.256	0.255	0.258	0.191	-	0.131
sign d.	0.240	0.240	0.243	0.165	0.131	-

Table 1: Root mean square errors among the randomization, t-test, bootstrap, Wilcoxon, sign, and sign minimum difference tests on 11986 pairs of TREC runs. This subset of the 18820 pairs eliminates pairs for which all tests agree the p-value was < 0.0001.

4. RESULTS

In this section, we report the amount of agreement among the p-values produced by the various significance tests. If the significance tests agree with each other, there is little practical difference among the tests.

We computed the root mean square error between each test and each other test. The root mean square error (RMSE) is:

$$RMSE = \left[\frac{1}{N}\sum_{i}^{N}(E_i - O_i)^2\right]^{1/2}$$

where E_i is the estimated p-value given by one test and O_i is the other test's p-value.

Table 1 shows the RMSE for each of the tests on a subset of the TREC run pairs. We formed this subset by removing all pairs for which all tests agreed the p-value was < 0.0001. This eliminated 6834 pairs and reduced the number of pairs from 18820 to 11986. This subset eliminates pairs that are obviously very different from each other — so different that a statistical test would likely never be used.

The randomization, bootstrap, and t tests largely agree with each other. The RMSE between these three tests is approximately 0.01, which is an error of 20% for a p-value of 0.05. An IR researcher testing systems similar to the TREC ad-hoc runs would find no practical difference among these three tests. The Wilcoxon and sign tests do not agree with any of the other tests. The sign minimum difference test better agrees with the other tests but still is significantly different. Of note, the sign and sign minimum difference tests produce significantly different p-values.

We also looked at a subset of 2868 TREC run pairs where either the randomization, bootstrap, or t test produced a p-value between 0.01 and 0.1. For this subset, the RMSE between the randomization, t-test, and bootstrap tests averaged 0.006, which is only a 12% error for a 0.05 p-value. For this subset of runs, the RMSE between these three tests and the Wilcoxon decreased but was still a large error of approximately 0.06. The sign and sign minimum difference showed little improvement.

Figure 4 shows the relationship between the randomization, bootstrap, and t tests' p-values. A point is drawn for each pair of runs. Both the t-test and bootstrap ap-

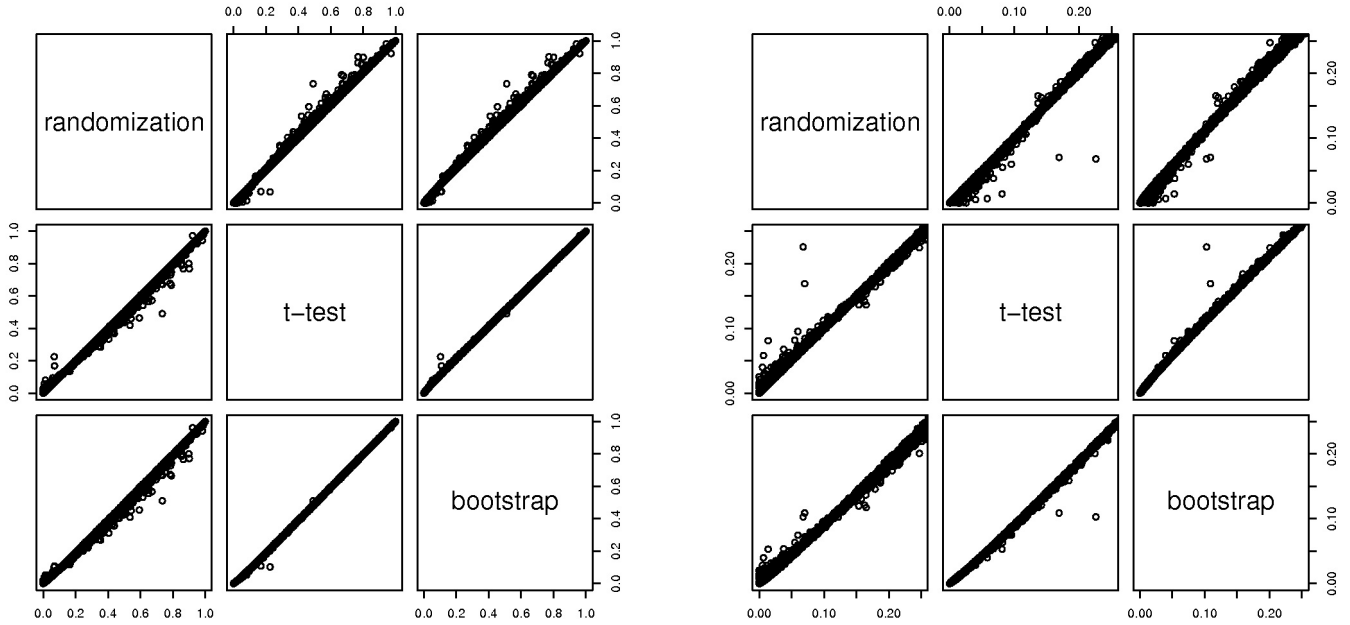

Figure 4: The three way agreement between the randomization test, the Student's t-test, and the bootstrap test. Plotted are the p-values for each test vs. each other test. The figure on the left shows the full range of p-values from 0 to 1 while the figure on the right shows a closer look at smaller p-values.

pear to have a tendency to be less confident in dissimilar pairs (small randomization test p-value) and produce larger p-values than the randomization test, but these tests find similar pairs to be more dissimilar than the randomization test. While the RMSE values in Table 1 say that overall the t-test agrees equally with the randomization and bootstrap tests, Figure 4 shows that the t-test has fewer outliers with the bootstrap.

Of note are two pairs of runs for which the randomization produces p-values of around 0.07, the bootstrap produces p-values of around 0.1, and the t-test produces much larger p-values (0.17 and 0.22). These two pairs may be rare examples of where the t-test's normality assumption leads to different p-values compared to the distribution free randomization and bootstrap tests.

Looking at the view of smaller p-values on the right of Figure 4, we see that the behavior between the three tests remains the same except that there is small but noticeable systematic bias towards smaller p-values for the bootstrap when compared to both the randomization and t tests. By adding 0.005 to the bootstrap p-values, we were able to reduce the overall RMSE between the bootstrap and t-test from 0.007 to 0.005 and from 0.011 to 0.009 for the randomization test.

Figure 5 plots the randomization test's p-value versus the Wilcoxon and sign minimum difference test's p-values. As variants of the randomization test, we use the randomization test for comparison purposes with these two tests. The different test statistics for the three tests leads to significantly different p-values. The bands for the sign test are a result of the limited number of p-values for the test. Compared to the randomization test, and thus to the t-test and bootstrap, the Wilcoxon and sign tests will result in failure to detect significance and false detections of significance.

5. DISCUSSION

To our understanding, the tests we evaluated are all valid tests. By valid, we mean that the test produces a p-value that is close to the theoretical p-value for the test statistic under the null hypothesis. Unless a researcher is inventing a new hypothesis test, an established test is not going to be wrong in and of itself.

A researcher may misapply a test by evaluating performance on one criterion and testing significance using a different criterion. For example, a researcher may decide to report a difference in the *median* average precision, but mistakenly test the significance of the difference in *mean* average precision. Or, the researcher may choose a test with an inappropriate null hypothesis.

The strong agreement among the randomization, t-test, and bootstrap shows that for the typical TREC style evaluation with 50 topics, there is no practical difference in the null hypotheses of these three tests.

Even though the Wilcoxon and sign tests have the same null hypothesis as the randomization test, these two tests utilize different criteria (test statistics) and produce very different p-values compared to all of the other tests.

The use of the sign and Wilcoxon tests should have ceased some time ago based simply on the fact that they test criteria that do not match the criteria of interest. The sign and Wilcoxon tests were appropriate before affordable computation, but are inappropriate today. The sign test retains

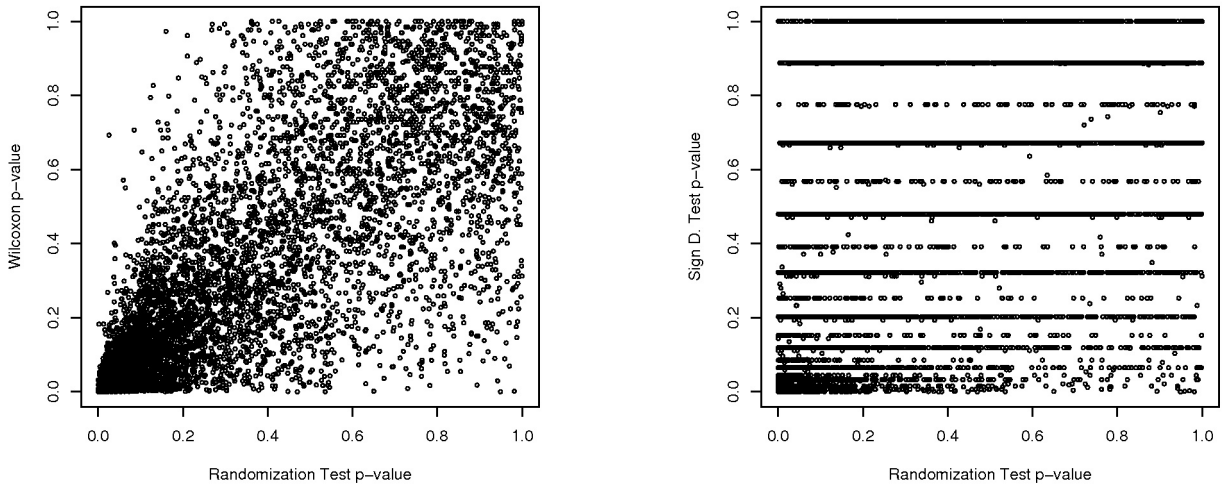

Figure 5: The relationship between the randomization test's p-values and the Wilcoxon and sign minimum difference tests' p-values. The Wilcoxon test is on the left and the sign minimum difference test is on the right. A point is drawn for each pair of runs. The x axis is the p-value produced by the randomization test run with 100K samples, and the y axis is the p-value of the other test.

validity if the only thing one can measure is a preference for one system over another and this preference has no scale, but for the majority of IR experiments, this scenario is not the case.

A researcher wanting a distribution-free test with no assumptions of random sampling should use the randomization test with the test statistic of their choice and not the Wilcoxon or sign tests.

5.1 Wilcoxon and Sign Tests

The Wilcoxon and sign tests are simplified variants of the randomization test. Both of these tests gained popularity before computer power made the randomization test feasible. Here we look at the degree to which use of these simplified tests results in errors compared to the randomization test.

Common practice is to declare results significant when a p-value is less than or equal to some value α. Often α is set to be 0.05 by researchers. It is somewhat misleading to turn the p-value into a binary decision. For example, there is little difference between a p-value of 0.049 and 0.051, but one is declared significant and the other not. Our preference is to report the p-value and flag results meeting the decision criteria.

Nevertheless, some decision must often be made between significant or not. Turning the p-value into a binary decision allows us to examine two questions about the comparative value of statistical tests:

1. What percent of significant results will a researcher mistakenly judge to be insignificant?
2. What percent of *reported* significant results will actually be insignificant?

We used a randomization test with 20 million samples to produce a highly accurate estimate of the p-value. Given its accuracy, we use it to judge which results are significant at various values of α for the null hypothesis of the randomization test. Recall that the null hypotheses of the Wilcoxon and sign tests are the same as the randomization test. The only difference between the randomization, Wilcoxon, and sign tests is that they have different test statistics. The randomization's test statistic matches our statistic of interest: the difference in mean average precision.

	Randomization Test	
Other Test	Significant	Not Significant
Significant	H = Hit	F = False Alarm
Not Significant	M = Miss	Z

Table 2: The randomization test is used to determine significance against some α. If the other test returns a p-value on the same side of α, it scores a hit or a correct rejection of the null hypothesis (Z). If the other test returns a p-value on the opposite side of α, it score a miss or a false alarm.

For example, if the randomization test estimates the p-value to be 0.006 and we set $\alpha = 0.01$, we will assume the result is significant. If another test estimates the p-value to be greater than α, that is a *miss*. If the other p-value is less than or equal to α, the other test scores a *hit*. When the randomization test finds the p-value to be greater than α, the other test can *false alarm* by returning a p-value less than α. Table 2 shows a contingency table summarizing hits, misses, and false alarms.

With these definitions of a hit, miss, and false alarm we can define the miss rate and false alarm ratio as measures

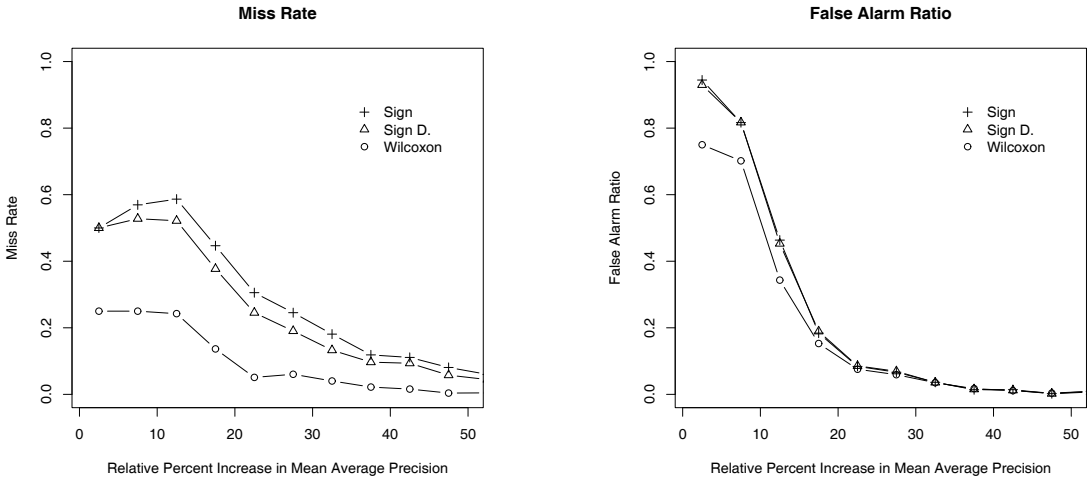

Figure 6: Miss rate and false alarm ratio for $\alpha = 0.1$.

of questions 1 and 2 above:

$$MissRate = \frac{M}{H + M}$$

where M is the number of misses and H is the number of hits.

$$FalseAlarmRatio = \frac{F}{H + F}$$

where F is the number of false alarms and H is the number of hits. The false alarm *ratio* is not the false alarm *rate*.

Another way to understand the questions we are addressing is as follows. A researcher is given access to two statistical significance tests. The researcher is told that one is much more accurate in its p-values. To get an understanding of how poor the poorer test is, the researcher says "I consider differences with p-values less than α to be significant. I always have. If I had used the better test instead of the poorer test, what percentage of my previously reported significant results would I now consider to be insignificant? On the flip side, how many significant results did I fail to publish?"

The miss rate and false alarm ratio can be thought of as the rates at which the researcher would be changing decisions of significance if the researcher switched from using the Wilcoxon or sign test and switched to the randomization test.

As we stated in the introduction, the goal of the researcher is to make progress by finding new methods that are better than existing methods and avoid the promotion of methods that are worse.

Figures 6 and 7 show the miss rate and false alarm ratio for the sign, sign minimum difference (sign d.), and Wilcoxon when α is set to 0.1 and 0.05. We show $\alpha = 0.1$ both as an "easy" significance level but also for the researcher who may be interested in the behavior of the tests when they produce one-sided p-values and $\alpha = 0.05$. In all cases, all of our tests produced two-sided p-values.

Given the ad-hoc TREC run pairs, if a researcher reports significance for a small improvement using the Wilcoxon or sign, we would have doubt in that result. Additionally, an IR researcher using the Wilcoxon or sign tests could fail to detect significant advances in IR techniques.

5.2 Randomization vs. Bootstrap vs. t-test

The randomization, bootstrap, and t tests all agreed with each other given the TREC runs. Which of these should one prefer to use over the others? One approach recommended by Hull [10] is to compute the p-value for all tests of interest and if they disagree look further at the experiment and the tests' criteria and null hypotheses to decide which test is most appropriate.

We have seen with the Wilcoxon and sign tests the mistakes an IR researcher can make using a significance test that utilizes one criterion while judging and presenting results using another criterion. This issue with the choice of test statistic goes beyond the Wilcoxon and sign tests. We ran an additional set of experiments where we calculated the p-value for the randomization test using the difference in median average precision. The p-values for the median do not agree with the p-values for the difference in mean average precision.

The IR researcher should select a significance test that uses the same test statistic as the researcher is using to compare systems. As a result, Student's t-test can only be used for the difference between means and not for the median or other test statistics. Both the randomization test and the bootstrap can be used with any test statistic.

While our experiment found little practical difference among the different null hypotheses of the randomization, bootstrap, and t tests, this may not always be so.

Researchers have been quite concerned that the null hypothesis of the t-test is not applicable to IR [19, 18, 21]. On our experimental data, this concern does not appear to be justified, but all of our experiments used a sample size N of 50 topics. $N = 50$ is a large sample. At smaller sample sizes, violations of normality may result in errors in the t-test. Cohen [4] makes the strong point that the randomization test performs as well as the t-test when the normality assumption is met but that the randomization test outperforms the t-test when the normality assumption is unmet. As such,

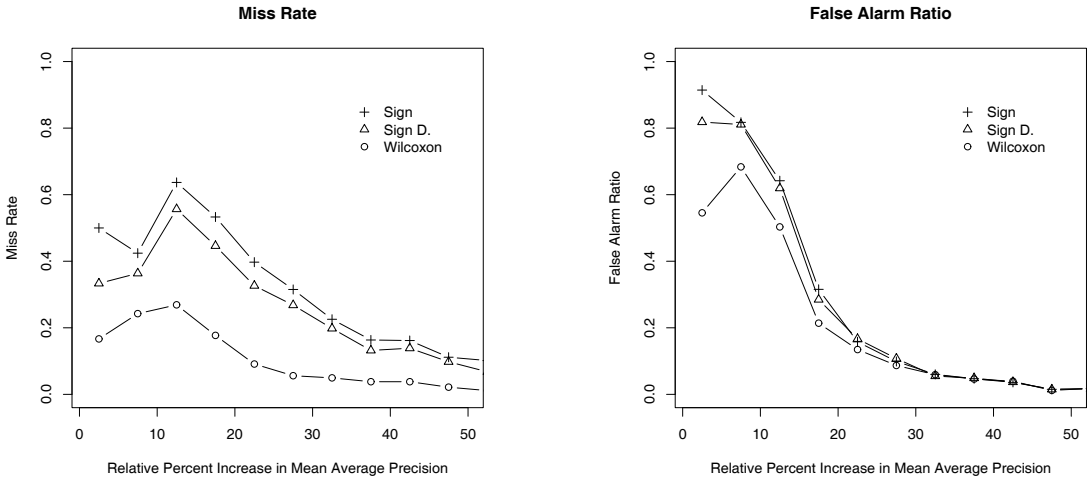

Figure 7: Miss rate and false alarm ratio for $\alpha = 0.05$.

the researcher is safe to use the randomization test in either case but must be wary of the t-test.

Between the randomization (permutation) test and the bootstrap, which is better? Efron invented the bootstrap in 1979. Efron and Tibshirani [8] write at the end of chapter 15:

> Permutation methods tend to apply to only a narrow range of problems. However when they apply, as in testing $F = G$ in a two-sample problem, they give gratifyingly exact answers without parametric assumptions. The bootstrap distribution was originally called the "combination distribution." It was designed to extend the virtues of permutation testing to the great majority of statistical problems where there is nothing to permute. When there *is* something to permute, as in Figure 15.1, it is a good idea to do so, even if other methods like the bootstrap are also brought to bear.

The randomization method does apply to the typical IR experiment. Noreen [14] has reservations about the use of the bootstrap for hypothesis testing.

Our largest concern with the bootstrap is the systematic bias towards smaller p-values we found in comparison to both the randomization and t tests. This bias may be an artifact of our implementation, but an issue with the bootstrap is the number of its possible variations and the need for expert guidance on its correct use. For example, a common technique is to *Studentize* the test statistic to improve the bootstrap's estimation of the p-value [8]. It is unclear when one needs to do this and additionally such a process would seem to limit the set of applicable test statistics. Unlike the bootstrap, the randomization test is simple to understand and implement.

Another issue with both the bootstrap and the t-test is that both of them have as part of their null hypotheses that the scores from the two IR systems are random samples from a single population. In contrast, the randomization test only concerns itself with the other possible experimental outcomes given the experimental data. The randomization test does not consider — the often incorrect — idea that the scores are random samples from a population.

The test topics used in TREC evaluations are not random samples from the population of topics. TREC topics are hand selected to meet various criteria such as the estimated number of relevant documents in the test collection [20]. Additionally, neither the assessors nor the document collection are random.

The randomization test looks only at the experiment and produces a probability that the experimental results could have occurred by chance without any assumption of random sampling from a population.

An IR researcher may argue that the assumption of random samples from a population is required to draw an inference from the experiment to the larger world. This cannot be the case. IR researchers have for long understood that inferences from their experiments must be carefully drawn given the construction of the test setup. Using a significance test based on the assumption of random sampling is not warranted for most IR research.

Given these fundamental difference between the randomization, bootstrap, and t tests, we recommend the randomization test be used when it is applicable. The randomization test is applicable to most IR experiments.

5.3 Other Metrics

Our results have focused on the mean average precision (MAP). We also looked at how the precision at 10 (P10), mean reciprocal rank (MRR), and R-precision affected the results. In general the tests behaved the same as for the MAP. Of note, the Wilcoxon test showed less variation for the MRR than for the other metrics.

6. RELATED WORK

Edgington's book [7] on randomization tests provides extensive coverage of the many aspects of the test and details how the test was created by Fisher in the 1930s and later was developed by many other statisticians. Box et al. pro-

vide an excellent explanation of the randomization test in chapter 4 of their classic text [1]. Efron and Tibshirani have a detailed chapter on the permutation (randomization) test in their book [8].

Kempthorne and Doerfler have shown that for a set of artificial distributions the randomization test is to be preferred to the Wilcoxon test which is to be preferred to the sign test [11]. In contrast, our analysis is based on the actual score distributions from IR retrieval systems.

Hull reviewed Student's t-test, the Wilcoxon signed rank test, and the sign test and stressed the value of significance testing in IR [10]. Hull's suggestion to compare the output of the tests was part of the inspiration for our experimental methodology. Hull also made the point that the t-test tends to be robust to violations of its normality assumption.

Wilbur compared the randomization, bootstrap, Wilcoxon, and sign tests for IR evaluation but excluded the t-test based on its normality assumption [21]. Wilbur found the randomization test and the bootstrap test to perform well, but recommended the bootstrap over the other tests in part because of its greater generality.

Savoy advocated the use of the bootstrap hypothesis test as a solution to the problem that the normality assumption required of the t-test is clearly violated by the score distributions of IR experiments [18]. Sakai used bootstrap significance tests to evaluate evaluation metrics [16], while our emphasis was on the comparison of significance tests.

Box et al. stress that when comparative experiments properly use randomization of test subjects, the t-test is usually robust to violations of its assumptions and can be used as an approximation to the randomization test [1]. We have confirmed this to be the case for IR score distributions.

Both Sanderson and Zobel [17] and Cormack and Lynam [5] have found that the t-test should be preferred to both the Wilcoxon and sign tests. We have taken the additional step of comparing these tests to the randomization and bootstrap tests that have been proposed by others for significance testing in IR evaluation.

7. CONCLUSION

For a large collection of TREC ad-hoc retrieval system pairs, the randomization test, the bootstrap shift method test, and Student's t-test all produce comparable significance values (p-values). Given that an IR researcher will obtain a similar p-value for each of these tests, there is no practical difference between them.

On the same set of experimental data, the Wilcoxon signed rank test and the sign test both produced very different p-values. These two tests are variants of the randomization test with different test statistics. Before affordable computation existed, both of these tests provided easy to compute, approximate levels of significance. In comparison to the randomization test, both the Wilcoxon and sign tests can incorrectly predict significance and can fail to detect significant results. IR researchers should discontinue use of the Wilcoxon and sign tests.

The t-test is only applicable for measuring the significance of the difference between means. Both the randomization and bootstrap tests can use test statistics other than the mean, e.g. the median. For IR evaluation, we recommend the use of the randomization test with a test statistic that matches the test statistic used to measure the difference between two systems.

8. ACKNOWLEDGMENTS

We thank Trevor Strohman for his helpful discussions and feedback on an earlier draft of this paper. We also thank the anonymous reviewers for their helpful comments.

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 HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. John Wiley & Sons, 1978.

[2] J. V. Bradley. *Distribution-Free Statistical Tests*. Prentice-Hall, 1968.

[3] C. Buckley. trec_eval. `http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz`.

[4] P. R. Cohen. *Empirical methods for artificial intelligence*. MIT Press, 1995.

[5] G. Cormack and T. Lynam. Validity and power of t-test for comparing map and gmap. In *SIGIR '07*. ACM Press, 2007.

[6] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *SIGIR '06*, pages 533–540. ACM Press, 2006.

[7] E. S. Edgington. *Randomization Tests*. Marcel Dekker, 1995.

[8] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1998.

[9] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, first edition, 1935.

[10] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93*, pages 329–338, New York, NY, USA, 1993. ACM Press.

[11] O. Kempthorne and T. E. Doerfler. The behavior of some significance tests under experimental randomization. *Biometrika*, 56(2):231–248, August 1969.

[12] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30, 1998.

[13] W. Mendenhall, D. D. Wackerly, and R. L. Scheaffer. *Mathematical Statistics with Applications*. PWS-KENT Publishing Company, 1990.

[14] E. W. Noreen. *Computer Intensive Methods for Testing Hypotheses*. John Wiley, 1989.

[15] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. 3-900051-07-0.

[16] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR '06*, pages 525–532. ACM Press, 2006.

[17] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05*, pages 162–169. ACM Press, 2005.

[18] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *IPM*, 33(4):495–512, 1997.

[19] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, second edition, 1979. `http://www.dcs.gla.ac.uk/Keith/Preface.html`.

[20] E. M. Voorhees and D. K. Harman, editors. *TREC*. MIT Press, 2005.

[21] W. J. Wilbur. Non-parametric significance tests of retrieval performance comparisons. *J. Inf. Sci.*, 20(4):270–284, 1994.

[22] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.