A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation

Julián Urbano[*]
jurbano@inf.uc3m.es

Mónica Marrero[*]
mmarrero@inf.uc3m.es

Diego Martín[†]
dmartin@dit.upm.es

[*]University Carlos III of Madrid
Department of Computer Science
Leganés, Spain

[†]Technical University of Madrid
Department of Telematics Engineering
Madrid, Spain

ABSTRACT

Previous research has suggested the permutation test as the theoretically optimal statistical significance test for IR evaluation, and advocated for the discontinuation of the Wilcoxon and sign tests. We present a large-scale study comprising nearly 60 million system comparisons showing that in practice the bootstrap, t-test and Wilcoxon test outperform the permutation test under different optimality criteria. We also show that actual error rates seem to be lower than the theoretically expected 5%, further confirming that we may actually be underestimating significance.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation.*

Keywords

Evaluation, Statistical significance, Randomization, Permutation, Bootstrap, Wilcoxon test, Student's t-test, Sign test.

1. INTRODUCTION

An Information Retrieval (IR) researcher is often faced with the question of which of two IR systems, A and B, performs better. She conducts an experiment with a test collection, and chooses an effectiveness measure such as Average Precision or nDCG. Based on the effectiveness difference she concludes that, for instance, system A is better. But we know there is inherent noise in the evaluation for a wealth of reasons concerning document collections, topic sets, relevance assessors, etc. Therefore the researcher needs the conclusion to be reliable, that is, the observed difference unlikely to have happened just by random chance. She employs a statistical significance test to compute this probability (the p-value). If $p \leq \alpha$ (the significance level, usually $\alpha = 0.05$ or $\alpha = 0.01$) the difference is considered statistically significant ($\mathsf{A} \succ\!\!\succ \mathsf{B}$). In practice this means that she can be confident that the difference measured with a similar test collection will be (at least) as large as currently observed. If $p > \alpha$ the difference is not significant ($\mathsf{A} \succ \mathsf{B}$), and she can not be confident that the observed difference is indeed real.

Unfortunately, there has been a debate regarding statistical significance testing in IR evaluation. Classical tests such as the paired t-test, the Wilcoxon test and the sign test make different assumptions about the distributions, and effectiveness scores from IR evaluations are known to violate these assumptions. The bootstrap test is an alternative that makes fewer assumptions and has other advantages over classical tests, and the permutation or randomization test is an even less stringent test in terms of assumptions that theoretically provides exact p-values. Because IR evaluations violate most of the assumptions, it is very important to know how robust these tests are in practice and which one is optimal.

Previous work [4, 5] compared these five tests with TREC Ad Hoc data, reaching the following conclusions: a) the bootstrap, t-test and permutation test largely agree with each other, so there is hardly any practical difference in using one or another; b) the permutation test should be the test of choice, though the t-test seems suitable as well; the bootstrap test shows a bias towards small p-values; c) the Wilcoxon and sign tests are unreliable and should be discontinued for IR evaluation. However, all these conclusions were based on the assumption that the permutation test is optimal. For example, authors showed that the Wilcoxon and sign tests fail to detect significance when the permutation test does and vice versa. That is, they are unreliable *according to* the permutation test.

But we may follow different criteria to chose an optimal test. We may want the test to be *powerful*, that is, to produce significant results as often as possible. Additionally, we may want it to be *safe* and yield low error rates so that it is unlikely that we draw wrong conclusions. But power and safety are inversely related; different tests show different relations depending on the significance level. The lower α the lower the power, because we need $p \leq \alpha$ for the result to be significant. Error rates are expected to be at the nominal α level, so the higher the significance level the higher the expected error rate. The test is *exact* if we can trust that the actual error rate is as dictated by the significance level. If it is below it means we are being too conservative and we are missing significant results; if it is above it means we are deeming as significant results that probably are not.

This paper presents a large-scale empirical study that compares all five tests according to these optimality criteria, providing significance and error rates at various significance levels for 50-topic sets. Our main findings are:

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

- In practice the bootstrap test is optimal in terms of power, the t-test is optimal in terms of safety, and the Wilcoxon test is optimal in terms of exactness.
- For all tests the actual error rate seems to be lower than the nominal 0.05 level, meaning that we are actually being too conservative.
- In practice the permutation test is not found to be optimal under any criterion.

2. DATA AND METHODS

To compare the five statistical significance tests at hand, we employed data from the TREC 2004 Robust Track. A total of 249 topics were used, 100 of which were originally developed in the TREC 7 and 8 Ad Hoc tracks (50 and 50). A total of 110 runs were submitted by 14 different groups. This dataset is unusually large both in terms of topics and runs, given that TREC tracks usually employ 50 topics. The subset with the 100 Ad Hoc topics is especially interesting: all 100 topics were developed and judged by the same assessors for the most part, and they were developed using the same methodology and pooling protocol with roughly the same number of runs contributing to the pools [6]. Additionally, all three tracks used disks 4 and 5 as document collection. Therefore, we can consider these two sets of 50 topics as two different samples drawn from the same universe of topics.

We randomly split these 100 topics into two disjoint subsets of 50 topics each: $\mathcal{T}$ and $\mathcal{T}'$. For each of these two subsets we evaluated all 110 runs as per Average Precision. This provides us with 5,995 system pairwise comparisons with $\mathcal{T}$ and another 5,995 with $\mathcal{T}'$. We ran all five statistical significance tests between each of these system pairs[1]. This gives us a total of 5,995 pairs of p-values per test, which can be regarded as the two p-values observed with two different test collections for any two systems. We performed 1,000 random trials of this experiment, so we have a total of 5,995,000 system pairwise comparisons and the corresponding 5,995,000 with another topic subset. Thus, this paper reports results on nearly 12 million p-values for each of the five tests, for a grand total of nearly 60 million p-values. To our knowledge, this is to date the largest study of this type.

Given an arbitrary topic set split, the 5,995 pairs of p-values provided by a test can be used to study its optimality. Consider a researcher that used topic subset $\mathcal{T}$ and ran a test to compute a p-value; under the significance level α he draws a conclusion. What can he expect with a different topic subset $\mathcal{T}'$?. One of these situations can occur:

- **Non-significance.** The result with $\mathcal{T}$ is $\mathsf{A} \succ \mathsf{B}$. We can really expect any result with $\mathcal{T}'$; there is a lack of statistical power in the experiment.
- **Success.** The result with both $\mathcal{T}$ and $\mathcal{T}'$ is $\mathsf{A} \succ\succ \mathsf{B}$. Both experiments show evidence of one system outperforming the other.
- **Lack of power.** The difference is $\mathsf{A} \succ\succ \mathsf{B}$ with $\mathcal{T}$ but it is $\mathsf{A} \succ \mathsf{B}$ with $\mathcal{T}'$. There is evidence of a lack of power in the second experiment.
- **Minor error.** The result with $\mathcal{T}$ is $\mathsf{A} \succ\succ \mathsf{B}$, but with $\mathcal{T}'$ it is $\mathsf{A} \prec \mathsf{B}$. The second experiment shows some evidence of a wrong conclusion in the first one.
- **Major error.** The result with $\mathcal{T}$ is $\mathsf{A} \succ\succ \mathsf{B}$, but with $\mathcal{T}'$ it is $\mathsf{A} \prec\prec \mathsf{B}$. The two experiments conflict.

[1]As in [4, 5], we calculated 100,000 samples in the permutation and bootstrap tests.

A powerful test minimizes the non-significance rate, a safe test minimizes the minor and major error rates, and an exact test maintains the global error rate at the nominal α level.

3. RESULTS

For every statistical significance test we computed the non-significance, success, lack of power and error rates at 32 significance levels $\alpha \in \{0.0001, ..., 0.0009, 0.001, ..., 0.009, ..., 0.1, ..., 0.5\}$. Tables 1 and 2 report the results for a selection of significance levels, and Figures 1 and 2 plot detailed views in the arguably most interesting $[0.001, 0.1]$ range. Please note that all plots are log-scaled.

Non-significance rate. The bootstrap test consistently produces smaller p-values, and it is therefore the most powerful of all tests across significance levels. Next are the permutation test for $\alpha < 0.01$ and the Wilcoxon test for the usual $\alpha \geq 0.01$. The t-test is consistently less powerful, though the difference is as small as roughly 1% fewer significant results at the usual $\alpha = 0.05$. The sign test is by far the least powerful of all five. Its stair-like behavior is explained by its resolution: p-values depend only on the sign of the score differences, not on the magnitude (see Figure 5 in [4]).

Success rate. The bootstrap and Wilcoxon tests are the most successful overall. For small significance levels $\alpha \leq 0.001$ the bootstrap test shows the highest success rate, but for the more usual levels $0.001 < \alpha \leq 0.05$ the Wilcoxon test performs better. Next are again the permutation test and the t-test, with very similar success rates about 0.3% lower than the Wilcoxon and bootstrap tests at the usual α levels. The sign test is clearly the worst of all.

Lack of power rate. Most of the unsuccessful comparisons are due to a lack of power with the second topic subset $\mathcal{T}'$. Relative results are comparable to results above: the bootstrap test dominates at small significance levels and the Wilcoxon tests dominates at the usual levels, again followed by the permutation test and the t-test.

Minor error rate. Except for rare occasions where the sign test's step-like behavior results in the smallest minor error rate, the t-test is generally the safest of all five across significance levels. The permutation test follows next with rates about 0.03% higher. The bootstrap test is consistently outperformed by the t-test and the permutation test; it yields 0.13% more minor errors at $\alpha = 0.05$. The Wilcoxon test performs even better than the permutation test for low significance levels, but it performs worse at the usual levels. As mentioned, the sign test wiggles between the other tests.

Major error rate. Similarly the t-test consistently performs best in terms of major errors, followed by the permutation and bootstrap tests. It is noticeable that for small significance levels neither of these three tests show any major error at all. For instance, at $\alpha = 0.005$ the t-test provides as many as 3,006,441 (50.2%) significant comparisons, and yet none of them results in a major error with the second topic subset. The Wilcoxon test outperforms the permutation test sporadically, but it performs worse overall. In general though, it is important to the bear in mind the magnitudes of the major error rates. For instance, at $\alpha = 0.05$ the t-test produced 1,082 major errors and the bootstrap test produced 1,523. While the difference may seem small compared to the total of significants (0.0277% vs 0.0383%), this is actually a large +41% relative increase. The sign test is clearly the worst of all, having an extremely large major error rate at small significance levels.

	Non-significance rate					Success rate					Lack of power rate				
α	t-test	perm.	boot.	Wilcox.	sign	t-test	perm.	boot.	Wilcox.	sign	t-test	perm.	boot.	Wilcox.	sign
.0001	.67698	.65006	**.6402**	.67189	.72367	.78749	.79451	**.79757**	.78859	.73691	.21222	.20503	**.20191**	.21107	.26264
.0005	.61184	.59202	**.58186**	.60471	.6782	.80788	.8107	**.8123**	.80765	.75479	.19138	.18827	**.18653**	.19146	.24431
.001	.5807	.56367	**.5532**	.5722	.63438	.81328	.81491	**.81547**	.8147	.76847	.18556	.18359	**.18285**	.18392	.22998
.005	.49842	.48755	**.47647**	.48911	.58347	.82051	.82145	.82365	**.82598**	.78018	.1764	.17503	.17243	**.17039**	.2169
.01	.45752	.44937	**.43847**	.4485	.53308	.82777	.82893	.83233	**.83338**	.79225	.16753	.16595	.16205	**.16115**	.20276
.05	.34779	.34539	**.33613**	.34215	.42762	.85579	.85565	.85856	**.85935**	.81999	.13157	.1314	.12743	**.12624**	.16709
.1	.29264	.29235	**.28412**	.28725	.37308	.86905	.86899	**.87086**	.86941	.83013	.1107	.11072	**.10736**	.10805	.15031
.5	.12398	.12581	.12153	**.11957**	.14934	.8836	.88369	.8836	**.88429**	.85641	.05175	.05232	.05088	**.04985**	.07511

Table 1: Non-significance rates over total of pairs (lower is better), success rates over total of significants (higher is better), and lack of power rates over total of significants (lower is better). Best per α in bold face.

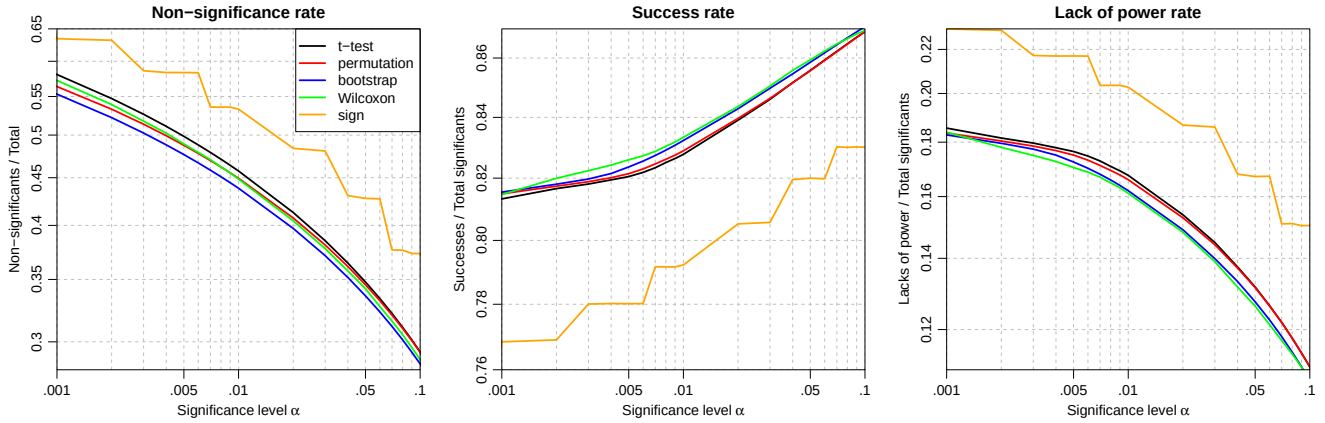

Figure 1: Non-significance rates over total of pairs (lower is better), success rates over total of significants (higher is better), and lack of power rates over total of significants (lower is better).

	Minor error rate					Major error rate					Global error rate				
α	t-test	perm.	boot.	Wilcox.	sign	t-test	perm.	boot.	Wilcox.	sign	t-test	perm.	boot.	Wilcox.	sign
.0001	**.00029**	.00046	.00051	.00034	.00045	**0**	**0**	**0**	5.08e-7	6.04e-7	**.00029**	.00046	.00051	.00034	.00045
.0005	**.00074**	.00104	.00117	.00089	.00089	**0**	**0**	**0**	4.22e-7	5.18e-7	**.00074**	.00104	.00117	.00089	.00089
.001	**.00116**	.00149	.00168	.00138	.00155	**0**	**0**	**0**	3.9e-7	4.56e-7	**.00116**	.00149	.00168	.00138	.00155
.005	.00309	.00352	.00392	.00362	**.00282**	**0**	**0**	6.37e-7	1.96e-6	.0001	.00309	.00352	**.00392**	.00362	.00292
.01	**.00469**	.00511	.0056	.00546	.00484	**.00001**	.00001	.00002	.00001	.00016	.0047	.00512	**.00562**	.00547	.00499
.05	**.01236**	.01264	.01363	.014	.01251	**.00028**	.0003	.00038	.0004	.00041	.01264	.01294	.01402	**.01441**	.01292
.1	.01903	.01906	.02027	.02123	**.01862**	.00122	.00123	.00152	.00131	**.00095**	.02025	.02029	.02178	**.02254**	.01956
.5	.03403	.03409	**.03389**	.03645	.03518	.03062	.0299	.03163	**.02941**	.0333	.06465	.06399	.06552	.06586	**.06849**

Table 2: Minor error rates over total of significants (lower is better), major error rates over total of significants (lower is better), and global error rates over total of significants ($errors = \alpha$ is better). Best per α in bold face.

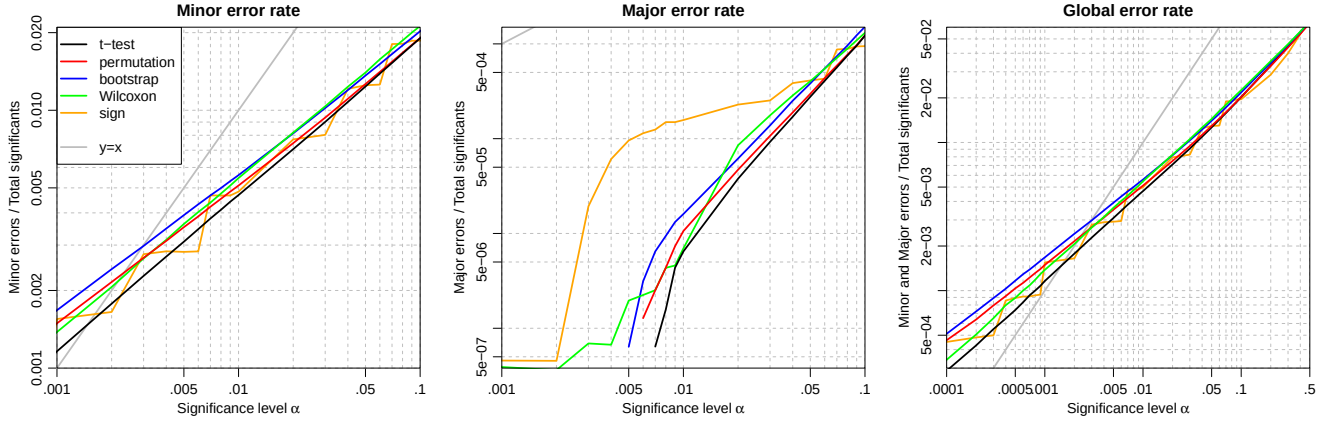

Figure 2: Minor error rates over total of significants (lower is better), major error rates over total of significants (lower is better), and global error rates over total of significants ($errors = \alpha$ is better).

Global error rate. Aggregating minor and major errors we have a global error rate that can be used as an overall indicator of test safety and exactness. Given the relative size of minor and major error rates, the trends are here nearly the same as fwith minor errors, but for the sake of completeness we plot the full range of significance levels. The t-test approximates best the nominal error rate for low significance levels, but the Wilcoxon test does better for the usual levels and best overall. Surprisingly the permutation test does not seem to be the most exact at any significance level.

4. DISCUSSION

Zobel [7] compared the t-test, Wilcoxon test and ANOVA at $\alpha = 0.05$, though with only one random split in 25-25 topics. He found lower error rates with the t-test than with the Wilcoxon test, and generally lower than the nominal 0.05 level. Given that the latter showed higher power and has more relaxed assumptions, he recommended it over the t-test. Sanderson and Zobel [3] ran a larger study also with splits of up to 25-25 topics. They found that the sign test has higher error rates than the Wilcoxon test, which has itself higher error rates than the t-test. They also suggested that the actual error rate is below the nominal 0.05 level when using 50 topic sets. Voorhees [6] also observed error rates below the nominal 0.05 level for the t-test, but more unstable effectiveness measures resulted in higher rates. Cormack and Lynam [1] used 124-124 topic splits and various significance levels. They found the Wilcoxon test more powerful than the t-test and sign test; and the t-test safer than the Wilcoxon and sign test. Sakai [2] proposed the bootstrap method for IR evaluation, but did not compare it with other tests.

Smucker et al. [4] compared the same five tests we study in this paper, arguing that the t-test, permutation and bootstrap tests largely agree with each other. Nonetheless, they report RMS Errors among their p-values of roughly 0.01, which is a large 20% for p-values of 0.05. Based on the argument that the permutation test is theoretically exact, they concluded that the Wilcoxon and sign tests are unreliable, suggesting that they should be discontinued for IR evaluation. They find the bootstrap test to be overly powerful, and given the appealing theoretical optimality of the permutation test they propose its use over the others, though the t-test admittedly performed very similarly. In a later paper [5] they found that the tests tended to disagree with smaller topic sets, though the t-test still showed acceptable agreement with the permutation test, again assumed to be optimal. The bootstrap test tended again to produce smaller p-values, so authors recommend caution if using it.

In this paper we ran a large-scale study to revisit these issues under different optimality criteria. In terms of safety, the t-test produced the smallest error rates across significance levels, followed by the Wilcoxon test for low levels and the permutation test for usual levels. In general, all tests yielded error rates higher than expected for low significance levels, but much lower for the usual levels. This suggests that we are being too conservative when assessing statistical significance at $\alpha = 0.05$; we expect 5% of our significant results to be wrong, but in practice only about 1.3% do indeed seem wrong. We must note though that this global error rate, as the sum of minor and major errors, is just an *approximation* of the true Type I error rates [1].

Table 3 shows the agreement of the five tests with themselves: p-values with topic subset $\mathcal{T}$ compared to those with subset $\mathcal{T}'$. The Wilcoxon test turns out to be the most stable of all for very small p-values, and generally more so than the permutation test. The t-test is the most stable overall. Indeed, if we compute the difference between the actual and nominal error rates we find that the Wilcoxon test is the one that best tracks the significance level and therefore seems to be the most exact (RMSE 0.1146), followed by the bootstrap, t-test, sign and permutation tests (RMSEs 0.1148, 0.1153, 0.1153 and 0.1155). This is particularly interesting for the bootstrap test: it provides the most significant results and the actual error rate is still lower than expected.

	t-test	perm.	boot.	Wilcox.	sign
$p \leq .0001$	.03603	.04348	.04475	**.03514**	.0556
$.0001 < p \leq .0005$	.10124	.11635	.11923	**.09976**	.13014
$.0005 < p \leq .001$	.13059	**.12999**	.1623	.14516	.14619
$.001 < p \leq .005$	**.16716**	.17044	.20032	.17841	.18024
$.005 < p \leq .01$	**.20724**	.21624	.2387	.21454	.21737
$.01 < p \leq .05$	**.25275**	.26685	.29801	.25779	.26114
$.05 < p \leq .1$	**.29734**	.31101	.33996	.30015	.30344
$.1 < p \leq .5$	**.31624**	.31855	.33816	.31804	.31802

Table 3: RMS Error of all five tests with themselves (lower is better). Best per bin in bold face.

In summary, a researcher that wants to maximize the number of significant results may use the more powerful bootstrap test and still be safe in the usual scenario. Researchers that want to maximize safety may use the t-test, and researchers that want to be able to trust the significance level may proceed with the Wilcoxon test. For large meta-analysis studies we encourage the use of the t-test and Wilcoxon test because they are far less computationally expensive and show near-optimal behavior. Unlike previous work concluded, our results suggest that in practice the permutation test is not optimal under any criterion. Further analysis with varied test collections and effectiveness measures should be conducted to clarify this matter, besides devising methods to better approximate what actual Type I error rates we have in IR evaluation. We further support the argument of discontinuing the sign test.

5. REFERENCES

[1] G. V. Cormack and T. R. Lynam. Validity and Power of t-test for Comparing MAP and GMAP. In *ACM SIGIR*, pages 753–754, 2007.

[2] T. Sakai. Evaluating Evaluation Metrics Based on the Bootstrap. In *ACM SIGIR*, pages 525–532, 2006.

[3] M. Sanderson and J. Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *ACM SIGIR*, pages 162–169, 2005.

[4] M. D. Smucker, J. Allan, and B. Carterette. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *ACM CIKM*, pages 623–632, 2007.

[5] M. D. Smucker, J. Allan, and B. Carterette. Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes. In *ACM SIGIR*, pages 630–631, 2009.

[6] E. M. Voorhees. Topic Set Size Redux. In *ACM SIGIR*, pages 806–807, 2009.

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