Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes

Mark D. Smucker
Department of Management Sciences
University of Waterloo
msmucker@uwaterloo.ca

James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
allan@cs.umass.edu

Ben Carterette
Department of Computer & Information Sciences
University of Delaware
carteret@cis.udel.edu

ABSTRACT

Research has shown that little practical difference exists between the randomization, Student's paired t, and bootstrap tests of statistical significance for TREC ad-hoc retrieval experiments with 50 topics. We compared these three tests on runs with topic sizes down to 10 topics. We found that these tests show increasing disagreement as the number of topics decreases. At smaller numbers of topics, the randomization test tended to produce smaller p-values than the t-test for p-values less than 0.1. The bootstrap exhibited a systematic bias towards p-values strictly less than the t-test with this bias increasing as the number of topics decreased. We recommend the use of the randomization test although the t-test appears to be suitable even when the number of topics is small.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Experimentation

Keywords: Statistical significance

1. INTRODUCTION

Information retrieval (IR) researchers rely on statistical significance tests to allow them to accurately detect and report significant improvements in performance. In an earlier work, we compared the randomization, bootstrap, Wilcoxon signed rank, sign, and Student's paired t tests of statistical significance as applied to IR evaluation [2]. By comparing the p-values produced by various statistical tests, one can determine if a practical difference exists between tests. For example, if two tests are in close agreement across different experiments, there is no practical difference in the tests to an IR researcher.

We found that the randomization, bootstrap, and t tests all largely agreed with each other while the Wilcoxon and sign tests disagreed with each other and the three other tests. Based on these results and the fundamental properties of the tests, we recommended the use of the randomization test but noted that if the test statistic of concern was the mean (as opposed to the median, e.g.) then the t-test appeared to be safe and robust to violations of the normality assumption.

Our earlier comparison only looked at TREC runs with 50 topics each. While 50 topics is standard for many of the TREC datasets, and the current push is for even larger numbers of topics [1], there are still valuable datasets with fewer topics. Moreover, even when large numbers of topics are available, IR researchers may want to perform analyses on smaller subsets of topics.

As the number of topics decreases, the possibility exists that the randomization, bootstrap, and t tests may cease to agree with each other. In particular, the t-test's robustness to violations of normality might not hold as the number of samples (topics) becomes small.

To test this, we compared the p-values produced by the randomization, bootstrap, and t tests for sample sizes of 10, 20, 30, 40, and 50 topics. We found that:

- Overall, the randomization, bootstrap, and t tests agree with each other but this agreement decreases as the number of topics decreases.
- The bootstrap tracks the t-test closely but with a systematic bias to produce smaller p-values.
- Even with a small number of topics, the t-test appears to be an acceptable significance test of the difference in mean performance for IR experiments.

2. METHODS AND MATERIALS

For each of the 18820 pairs of the ad-hoc retrieval runs of TREC 3, 5–8, we computed the two-sided statistical significance (p-value) of the difference in the pair's mean average precision using each of three tests: the randomization, shifted bootstrap, and Student's paired t-test. Both the randomization and bootstrap are distribution-free tests. Space limitations prevent us from explaining the details of each of these well-known tests.

For both the randomization and bootstrap, we performed 100,000 samples. For each pair of runs, we sampled topics without replacement to produce runs with 10, 20, 30, and 40 topics. To compare significance tests, we computed the root mean square error between each test and each other test's p-values. The root mean square error is:

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

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Pairs of TREC runs with p-values ≥ 0.0001

	Number of Topics				
	50	40	30	20	10
rand. vs. t-test	0.007	0.009	0.011	0.018	0.037
boot. vs. t-test	0.007	0.009	0.011	0.017	0.035
boot. vs. rand.	0.011	0.014	0.017	0.026	0.051

Run pairs with p-value p such that $0.0001 < p < 0.5$

	Number of Topics				
	50	40	30	20	10
rand. vs. t-test	0.005	0.006	0.008	0.012	0.027
boot. vs. t-test	0.008	0.010	0.013	0.020	0.041
boot. vs. rand.	0.010	0.013	0.016	0.024	0.047

Table 1: The root mean square error among the randomization (rand.), t-test, and the bootstrap (boot.) test's p-values for pairs of TREC runs such that all three tests agree that the p-value p is $\geq$ 0.0001 (top) and $0.0001 < p < 0.5$ (bottom).

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

3. RESULTS AND DISCUSSION

Table 1 (top) shows the root mean square error (RMSE) between the three tests for different numbers of topics. These results show that all three tests largely agree with each other but as the sample size (number of topics) decreases, the agreement decreases. In line with the results found for 50 topics, the randomization and bootstrap tests agree more with the t-test than with each other.

We looked at pairwise scatterplots of the three tests at the different topic sizes. While there is some disagreement among the tests at large p-values, i.e. those greater than 0.5, none of the tests would predict such a run pair to have a significant difference. More interesting to us is the behavior of the tests for run pairs with lower p-values.

Table 1 (bottom) shows the RMSE among the three tests for run pairs that all three tests agreed had a p-value greater than 0.0001 and less than 0.5. In contrast to all pairs with p-values ≥ 0.0001 (Table 1 top), these run pairs are of more importance to the IR researcher since they are the runs that require a statistical test to judge the significance of the performance difference. For these run pairs, the randomization and t tests are much more in agreement with each other than the bootstrap is with either of the other two tests.

Looking at scatterplots, we found that the bootstrap tracks the t-test very well but shows a systematic bias to produce p-values smaller than the t-test. As the number of topics decreases, this bias becomes more pronounced. Figure 1 shows a pairwise scatterplot of the three tests when the number of topics is 10. The randomization test also tends to produce smaller p-values than the t-test for run pairs where the t-test estimated a p-value smaller than 0.1, but at the same time, produces some p-values greater than the t-test's. As Figure 1 shows, the bootstrap consistently gives smaller p-values than the t-test for these smaller p-values.

While the bootstrap and the randomization test disagree with each other more than with the t-test, Figure 1 shows that for a low number of topics, the randomization test shows less noise in its agreement with the bootstrap compared to the t-test for small p-values.

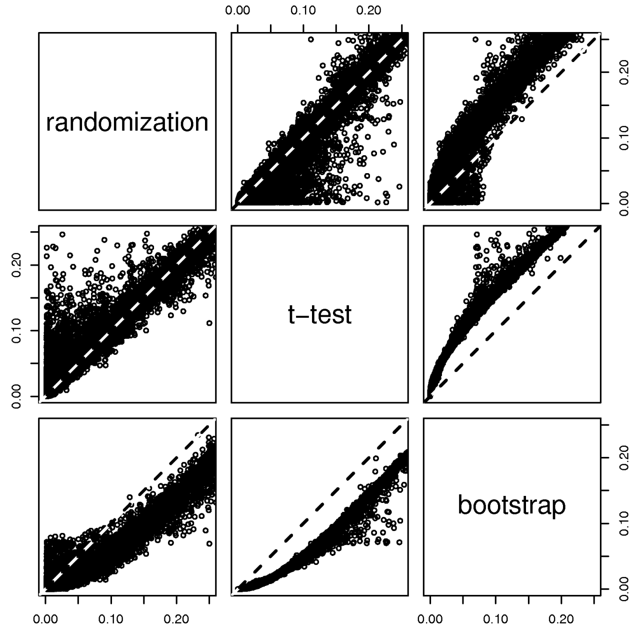

Figure 1: A pairwise comparison of the p-values less than 0.25 produced by the randomization, t-test, and the bootstrap tests for pairs of TREC runs with only 10 topics. The small number of topics highlights the differences between the three tests.

4. CONCLUSION

Using a large collection of TREC retrieval experiments, we compared the p-values produced by three tests of statistical significance: randomization, bootstrap, and Student's paired t-test, across different sample sizes (number of TREC topics). Overall, the three tests agree with each other, but the agreement among the tests decreases as the sample size decreases.

We found little to no evidence to reject the t-test, but the bootstrap looks suspicious with its bias to produce smaller p-values. Thus, if an IR researcher wants a distribution-free test or uses a test statistic other than the mean, we recommend the randomization test. The t-test appears suitable even for smaller sample sizes when the test statistic is the mean.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, and in part by the University of Waterloo. 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.

6. REFERENCES

[1] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *SIGIR '08*, pages 651–658. ACM Press, 2008.

[2] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07*, pages 623–632. ACM Press, 2007.