Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation

Javier Parapar
Information Retrieval Lab
Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC)
Universidade da Coruña
A Coruña, Spain
javier.parapar@udc.es

David E. Losada
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
david.losada@usc.es

Álvaro Barreiro
Information Retrieval Lab
Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC)
Universidade da Coruña
A Coruña, Spain
alvaro.barreiro@udc.es

ABSTRACT

Null Hypothesis Significance Testing (NHST) has been recurrently employed as the reference framework to assess the difference in performance between Information Retrieval (IR) systems. IR practitioners customarily apply significance tests, such as the t-test, the Wilcoxon Signed Rank test, the Permutation test, the Sign test or the Bootstrap test. However, the question of which of these tests is the most reliable in IR experimentation is still controversial. Different authors have tried to shed light on this issue, but their conclusions are not in agreement. In this paper, we present a new methodology for assessing the behavior of significance tests in typical ranking tasks. Our method creates models from the search systems and uses those models to simulate different inputs to the significance tests. With such an approach, we can control the experimental conditions and run experiments with full knowledge about the truth or falseness of the null hypothesis. Following our methodology, we computed a series of simulations that estimate the proportion of Type I and Type II errors made by different tests. Results conclusively suggest that the Wilcoxon test is the most reliable test and, thus, IR practitioners should adopt it as the reference tool to assess differences between IR systems.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

information retrieval, statistical testing, simulation

ACM Reference Format:
Javier Parapar, David E. Losada, and Álvaro Barreiro. 2021. Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC'21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3412841.3441945

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SAC '21, March 22–26, 2021, Virtual Event,Republic of Korea
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8104-8/21/03…$15.00
https://doi.org/10.1145/3412841.3441945

1 INTRODUCTION

In this paper, we propose a fitting approach to model the pattern of appearance of relevant documents in ranked sets of search results. The resulting models can simulate the output of search systems and, thus, they can support several meaningful tasks. We focus here on introducing the learned models as useful devices for studying tests of statistical significance.

Evaluation is essential to advance in building better Information Retrieval (IR) algorithms. Search experiments produce performance data, and significance tests are required to decide whether the experiments show that there is a statistically meaningful difference between retrieval methods. Significance tests pervade IR experimentation, but there is still no agreement on which test IR practitioners should apply.

Recent papers [11, 21] have manifested the limitations of query-splitting and permutation-based studies for comparing significance tests for IR. These studies have, however, either relied on the existence of Score Distribution models or worked with fits of effectiveness scores whose quality is unknown (and these fits alter the experimental conditions in a way that favors some test over the others). Our innovative modeling approach permits us to compare significance tests based on simulating how search systems retrieve relevant documents in the rankings. The method does not require retrieval scores, and we provide evidence on the validity of our simulation of rankings.

Score Distributions (SDs) are useful devices in several tasks, but the dependence on score distribution modeling is limiting. SD models depend on the availability of actual scores generated by systems and previous studies that have employed SDs when simulating search systems to evaluate significance tests, e.g., [11], could only evaluate systems that provide retrieval scores. This is an important limitation because many search systems (even in TREC) do not reveal document retrieval scores. Furthermore, as argued in [13], SD models often make strong assumptions about the retrieval scores. The work presented here complements recent work on evaluating significance tests for IR experiments [11, 21]. Our approach does not require SDs and can, therefore, be employed to run simulations of any search system that provides a rank of documents by decreasing estimated relevance. In this way, our comparison of significance tests is more exhaustive than previous experiments because it models a broader range of retrieval systems. Moreover, our approach does not model the distribution of any effectiveness

score and, thus, our simulation method is agnostic to the type of distribution followed by the effectiveness metrics.

A core strength of our comparison is that, inspired by [11], we assess significance tests under perfect knowledge about the truth or falseness of the null hypothesis. Using simulated search systems, we can produce two rankings for the same system-query pair (null hypothesis true), or we can manipulate the parameters and produce rankings from simulated models whose performance is increasingly different (null hypothesis false). With such certainty on the source of the rankings (same model producing two rankings vs. two different models producing two rankings), we can accurately evaluate the type I and type II errors made by the significance tests. This robust evaluation mechanism, together with the ability to model and simulate any search system (even those that do not reveal actual scores of documents), are the primary building blocks of our framework to compare significance tests in IR.

The rest of the paper is organized as follows. Section 2 provides the context for our experiments by reviewing appropriate literature. The simulation approach is presented and validated in Section 3. The experiments performed to evaluate significance tests are reported in Section 4, and conclusions are drawn in Section 5.

2 RELATED WORK

Parapar and colleagues [11] recently published a novel proposal to evaluate significance tests based on explicit knowledge about the null hypothesis. This comparison is valuable, but it relies on the existence of retrieval scores. The authors modeled IR systems using Score Distributions [9] learned from TREC runs, produced simulated search results from such models, and compared them using various significance tests. The use of Score Distribution models limits the simulation to those TREC systems that revealed document scores. In this work, we investigate whether or not the relative merits of the significance tests held in the absence of retrieval scores. This type of study is essential because most retrieval experimentations (e.g., in TREC or CLEF) need to compare systems that do not always supply retrieval scores.

A later study by Urbano and colleagues [21], also proposed a simulation based on explicit knowledge about the null hypothesis for evaluating significance tests. Their method does not model the production of rankings but, instead, works from the individual effectiveness scores of a system over a set of queries (e.g., a sequence of Average Precision –AP– values). This sequence of scores is used to fit a distribution and, next, pseudo-observations are drawn and fed to a copula-based method that models the dependence among systems. A fundamental limitation of this approach is that the fits are done with specific families of distributions[1] and, as argued in [22], the results "do not tell us ... themselves about the quality of the fitted models". If simulated models are fitted from pre-selected classes of distributions, the comparison is biased towards significance tests that follow certain parametric assumptions. For example, the t-test profits from simulated data that follows normal-shaped distributions (see discussion in Section 4.2, Figure 6). In summary, such pre-selection of certain parametric distributions is an artifact of the simulation proposed in [21], and the best fit for each combination of measure and retrieval system might be still a poor fit. Furthermore, such simulation approach is inherently tied to specific performance measures, while the method proposed here produces simulated rankings of relevant and non-relevant documents and we can compute any performance metric from the resulting simulations.

Besides these two recent studies, there are several papers in the literature that compared and analyzed significance tests in IR. Research works in this area are grouped into two main classes: query-splitting and permutation-based. Studies in the first group divide the available query set into two disjoint subsets and, subsequently, study the behavior of the significance tests over the two query splits. Following this approach, Zobel [24] concluded that the Wilcoxon test has more power and is more reliable than both the t-test and ANOVA. In his study, a type I error was recorded when a statistically significant difference between two systems was observed in the first query split, and the ordering of the systems was different in the second query split. Following a similar query-splitting approach, Sanderson and Zobel [18] later suggested that the t-test has lower error rates when compared to the sign test and the Wilcoxon test, and Cormack and Lynam [7] found that the t-test, Wilcoxon and the sign test have high power and are reliable. Although these investigations provided valuable insights for IR evaluation, their conclusions should be interpreted with care. The query-splitting methodology works with arbitrary (and small) splits of queries and lacks knowledge about the truth of the null hypothesis. A given significance test might give consistent results over two query splits, but this does not mean that the result of the test is correct. The test might be consistently rejecting a null hypothesis that is true or, conversely, it might be consistently accepting a null hypothesis that is actually false. Note also that, for any significance test, the smaller the sample size, the lower the power and, thus, by applying query splitting we are setting a barrier that limits the differences that can be detected by the tests.

Working with two-sample t-tests, Student's and Welch's, Sakai [16] followed a query-splitting approach that had restricted knowledge about the null hypothesis. Given a query set and the associated retrieval results produced by some systems, the queries were randomly split into two sets. For each system and evaluation metric, a two-sided, two-sample test was run to assess whether or not the difference between the two means of the same system were significant. Since the means came from the same system, this is a case where the null hypothesis is assumed to be true. This approach works with unpaired data and, thus, cannot be applied to compare significance tests in standard retrieval experiments (where all rankers evaluated with the same queries). This unpaired method is, however, valuable in situations with two-samples, such as comparisons of clickthrough data from two search engines.

The second class of comparisons of significance tests was founded by Smucker and colleagues [20], who employed the permutation test as the main reference to evaluate significance tests. More specifically, a false alarm was recorded when a significance test marked a difference as significant while the permutation test had labeled the same case as non-significant. Following such analysis, they suggested that the use of the Wilcoxon test and the sign test should be discontinued because these two tests show a discordant behavior

[1] Truncated Normal, Beta, Truncated Normal Kernel Smoothing, Beta Kernel Smoothing, Beta-Binomial and Discrete Kernel Smoothing.

when compared with the Bootstrap test, the t-test, and the permutation test. However, such a conclusion was based on taking the permutation's decisions as the correct ones. The permutation test computes good approximations to the actual p-values, but this does not mean that its decisions on statistical significance should be taken as golden truth labels. As a matter of fact, by design, the test makes an α proportion of type I errors. The experiments reported in our paper also show that the Wilcoxon test and the sign test are discordant with the other tests, but we demonstrate that this outcome derives from the fact that both of them have more power than the other tests compared. These results are in agreement with those found in recent SD-based simulations of retrieval systems [11].

3 BUILDING AND EVALUATING SIMULATED SEARCH SYSTEMS

An essential component of our approach is to model how retrieval systems produce a ranked list of relevant documents. For each query, typical IR collections (e.g., TREC) supply relevance judgments and rankings of documents produced by several participants. Given each query-system pair, we model the presence of relevant documents in the ranking produced by the system as follows. For each position p in the ranked list, we know the relevance value ($r_p \in \{0, 1\}$) of the document ranked at the p-th position[2]. With logistic regression, we can model how relevance decays as we go down in the rankings. We build a logistic regression model where the target variable is relevance, the only predictor is the position in the rank, and the training set contains as many examples as ranked documents: $\{(1, r_1), \ldots, (ranksize, r_{ranksize})\}$. The fitted model has the form:

$$h_\theta(p) \quad = \quad \frac{1}{1 + e^{-\theta_0 - \theta_1 \cdot p}} \tag{1}$$

Figure 1 shows an example of this fit for one pair (system, query). This model can be employed to simulate multiple rankings for the same system-query pair. To meet this aim, we produce a new ranking by iteratively drawing Bernoulli samples over each position. The Bernoulli distribution at each position is defined from the fitted Logistic model at that position. Algorithm 1 illustrates this process. Observe that this simulation approach does not produce actual ranked documents but relevance values (the variable *Ranking* of size *rank_size* stores a sequence of 0s and 1s). This sequence suffices to compute any IR performance metric and, next, evaluate the significance tests.

Algorithm 1: Algorithm for simulating a new ranking.

Input: A logistic regression fitted model, h_θ

```
Ranking ←{};
for position ← 1 ... rank_size do
    BernoulliParam ← h_θ(position);
    Draw a sample rel_position ~ Bern(BernoulliParam);
    Ranking[position] ← rel_position
end
```

[2]In this paper we work with binary relevance, but these models are potentially applicable to fit systems with graded relevance judgments.

To compare significance tests under H0 (null hypothesis true), we obtain the fits of the system (e.g., 50 fits for a typical TREC collection with 50 queries), we produce two simulations from the same system (two new rankings for each fit), we compute the associated performance metric (e.g., yielding 50 Average Precision values for each simulation) and, next, we input these two sequences of performance values to the significance tests. By repeating this process over multiple systems and repetitions, we can evaluate how effective significance tests are. This experiment allows us to estimate the probability of a type I error, $P(Reject\ H0|H0)$.

By manipulating the parameters of the logistic regression model, θ_0 and θ_1, we can also simulate the situation where H0 is false. Given a fit obtained for a certain system-query pair, we can obtain a better or worse retrieval system by producing a new logistic regressor whose probability of relevance (h_θ) is higher or lower, respectively. To this aim, we increase or decrease the values of θ_0 and θ_1 by a given proportion. The probability of relevance grows with $\theta_0 + \theta_1 \cdot p$ and, thus, we produce an improved version of the system by setting $\theta_i^* = \theta_i \cdot (1 + prop)^s$, where s equals 1 if θ_i is positive and equals -1 if θ_i is negative, and $prop$ is the proportion parameter. For example, with $prop = 0.1$ (10%) and $\theta_1 = -0.2$ this leads to $\theta_1^* = -0.2 \cdot (1.1)^{-1} = -0.18$. Figure 2a shows that the average MAP of the systems grows with $prop$. This effect is a natural consequence of the higher production of relevant documents in the rankings. In our experiments, we obtain the original fits of a system, and we produce a better model for each fit. Observe that the simulation is stochastic and, thus, the better versions of the individual query models do not always produce better performance than the original models. Figure 2b plots the effect of a 5% manipulation of the parameters. This boxplot demonstrates that the resulting model is generally more effective but some queries are improved while other queries are deteriorated. Given the two sequences of performance metrics (e.g. APs from original model vs APs from the improved model), we input them to the significance tests and study their ability to detect the difference. If a given significance test does not reject the null hypothesis then we record a type II error (non-rejection of a false null hypothesis). By doing this experiment with increasingly higher differences between the two models (higher $prop$, i.e. higher effect size), we can analyze the power of the tests under different effect sizes. As argued in [14], the output of the significance tests should be considered in combination with the effect size.

Table 1 reports the main statistics of the collections used for experimentation. The first experiment aimed to validate the simulated models. To demonstrate that the fitted models are good representatives of the original models, we ranked all the original TREC systems with the original qrels and compared this ranking against a ranking of systems produced by simulation. The original ranking represents the relative ordering of the real systems, while the second ranking represents the relative ordering of systems that are simulated from the original ones. We used Mean Average Precision (MAP) as the reference metric for ordering systems, and we ran 1000 simulations for each system. Table 2 shows the average correlations between original ranking and simulated rankings. These correlation figures (and the associated p-values) demonstrate that the simulation does not significantly alter the ranking of systems, which is an integral part of IR evaluation. As argued by Voorhees

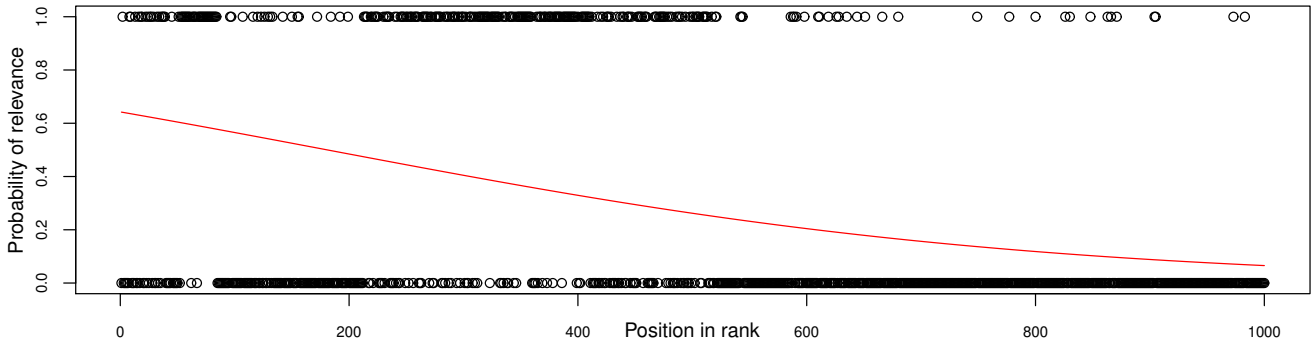

Figure 1: Fit obtained from one TREC5 system for query 273. The X-axis represents the rank positions, and the Y-axis represents the probability of finding a relevant document at each position. The circles in the plot are the relevant documents (rel=1) or non-relevant documents (rel=0) retrieved by the system. The solid line is the fitted curve estimated with logistic regression.

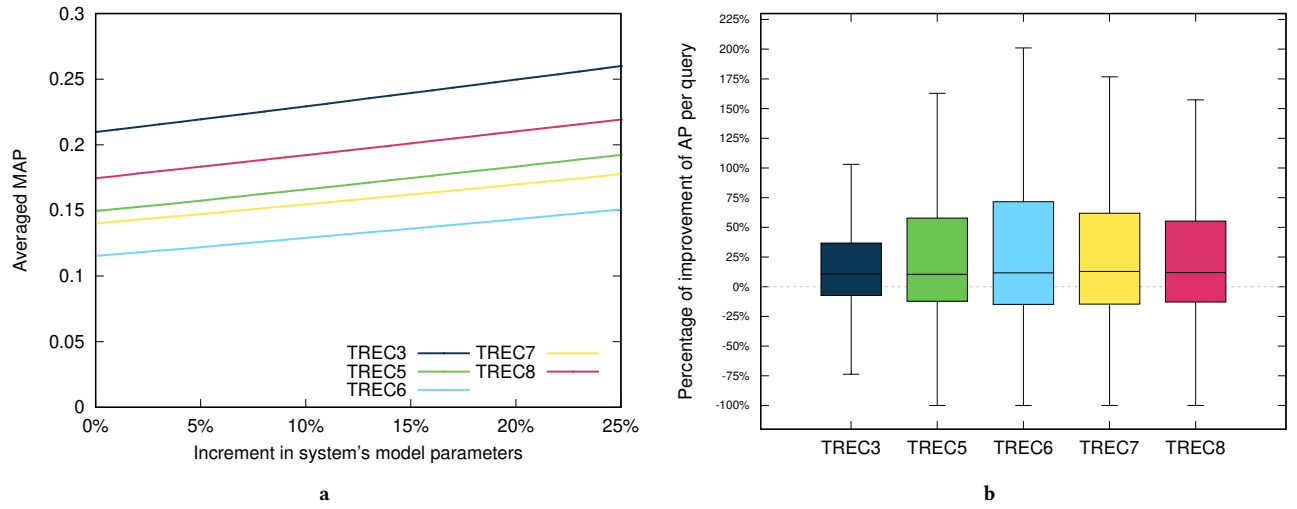

Figure 2: Validation experiments: (a) Average MAP against increasing improvements of the systems. (b) Effect of a 5% improvement in the systems (for each collection, the boxplot summarises the distribution of improvements across queries in 100 repetitions of each query for all queries and participant systems).

[23], levels of correlation above .85 indicate that the compared rankings are highly similar and, although not identical, the simulated rankings do not impact noticeably on the reliability of the comparison. Additionally, we tested the equality of the two AP distributions (actual TREC system vs simulated TREC system). To this aim, we ran the Cucconi non-parametric test of equality of distributions. This test has been shown to be robust to changes in the scale and location of the distribution [10]. We applied Cucconi's tests on the available TREC systems ($\alpha = 0.05$) and found no noticeable differences between the original and the simulated systems.

4 EXPERIMENTS

In our experimental study we employed the simulated systems to evaluate the following significance tests: the t-test, the Wilcoxon signed rank test, the sign test, the permutation test, and the Bootstrap test[3]. A full description of these tests can be found in [6] and brief explanations on the use of the tests in IR experiments are available in [11, 20]. We will focus our analysis on the two-sided paired-sample case, which naturally arises in standard IR experiments where two retrieval systems ran the same sets of queries. Average Precision was taken as effectiveness metric. Moreover, we also performed additional experiments with one-sided paired tests and other performance metrics (we briefly discuss those results in Section 4.3). Our simulations can also support the evaluation of tests oriented to multiple comparisons and repeated measures [17], but we left this comparison as future work.

[3]We experimented with the Bootstrap test as defined in [20].

Table 1: TREC collections and system runs used in our experiments (ad-hoc – Category A – tracks). The last two columns report the average and standard deviation of the MAPs of the participating systems.

Edition	Topics	# Relevant docs.	# Systems	μ MAP	σ MAP
TREC 3	151-200	9805	40	0.2573	0.0848
TREC 5	251-300	5524	61	0.1898	0.0676
TREC 6	301-350	4611	74	0.1716	0.0876
TREC 7	351-400	4674	103	0.1991	0.0802
TREC 8	401-450	4728	129	0.2345	0.0965

Table 2: Correlation between the original ranking of systems and the ranking of systems produced by simulation. The systems were ranked by decreasing MAP. The p-values for the significance testing (positive one-sided) of the correlation values are also reported.

	Pearson's r	p-value	Kendall's τ	p-value	Spearman ρ	p-value
TREC 3	0.9900	$3{,}1 \cdot 10^{-34}$	0.9077	$2{,}2 \cdot 10^{-16}$	0.9812	0,0
TREC 5	0.9649	$2{,}8 \cdot 10^{-36}$	0.8459	$2{,}9 \cdot 10^{-22}$	0.9621	0,0
TREC 6	0.9686	$1{,}4 \cdot 10^{-45}$	0.8534	$2{,}8 \cdot 10^{-27}$	0.9636	0,0
TREC 7	0.9476	$3{,}8 \cdot 10^{-52}$	0.8473	$3{,}7 \cdot 10^{-37}$	0.9604	0,0
TREC 8	0.9695	$1{,}1 \cdot 10^{-79}$	0.8216	$1{,}1 \cdot 10^{-43}$	0.9479	$3{,}1 \cdot 10^{-65}$

The experiments are fully reproducible, and the code that builds the models and runs the simulations is available at our institutional website[4].

4.1 Null hypothesis true

Figure 3 summarizes the results obtained when the null hypothesis is true (the compared samples are taken from the same simulated system without modifying the *prop* parameter). For each TREC collection, the results correspond with the comparison of all available systems against themselves over 10,000 repetitions. Experiments were done with three configurations of query sizes (10, 30 and 50 queries, chosen randomly from the available queries). Given a significance value, the estimation of type I error is obtained by computing the proportion of comparisons where the significance test rejected H0.

By design, significance tests are expected to have a proportion of type I errors that matches the significance level (set to 0.05 in Figure 3). For each collection and significance test, the bars show that the higher number of queries the closer the test gets to the expected proportion of type I errors. This is in accordance with the long-held recommendation that IR experiments should not have a low number of queries.

Wilcoxon and permutation tests achieve the expected proportion of type I errors. The other tests show a lower proportion of this type of errors. From a practical perspective, these lower counts of errors might seem convenient but this outcome shows that the p-values obtained by these three tests do not accurately estimate the probability of finding the observed difference between systems when H0 is true. The study reported in [11] only experimented with systems that reveal retrieval scores but also found that Wilcoxon and permutation both match the expected ratio of type I errors. To further analyse this point, we ran additional experiments with 50 queries, 10,000 repetitions and varying significance values[5]. Figure 4 shows the results of this experiment. Again, Wilcoxon and permutation show a pattern that fits with the target proportion of errors, while the other tests (particularly, Bootstrap) show significant deviations.

4.2 Null hypothesis false

Next, we ran a series of experiments where each model was compared against an improved variant of the same model. These experiments are reported in Fig.5. The lefmost point (0%) corresponds with the case of equal models, while the rest of the points correspond with increasingly larger differences between the models being compared (i.e., increased effect size by varying $prop \in \{0.005, 0.010, 0.015, ..., 0.250\}$). For every TREC edition and effect size, each system was compared against its improved version (and this comparison was repeated 10,000 times). For each comparison, a sample of 50 AP values was drawn from the 50 original models and another sample of 50 AP values was drawn from the altered models[6].

The Wilcoxon test is a clear winner. Under all circumstances, it performs better at rejecting the null hypothesis. The fact that the Wilcoxon test has higher power than the other tests was also shown in [11], and this finding is consistent with authorative studies on Statistics. For example, Conover [6] demonstrated that the t-test has less power than permutation and permutation has less power than Wilcoxon.

The Wilcoxon test and the sign test make fewer assumptions about the data. Other tests, such as the t-test, use more information from the data in their statistics (e.g., magnitudes of the differences) and one could think that this would be an advantage to detect a difference. Effectiveness data, however, rarely satisfy the conditions imposed by parametric tests. The t-test, for instance, assumes normality on the data. In Figure 6 (left) we can observe how the

[4] https://www.dc.fi.udc.es/~parapar/testing-tests

[5] $\alpha \in \{0.001, 0.002, ..., 0.009, 0.01, 0.02, ..., 0.09, 0.1\}$

[6] Note that each system-query pair produces an individual fit.

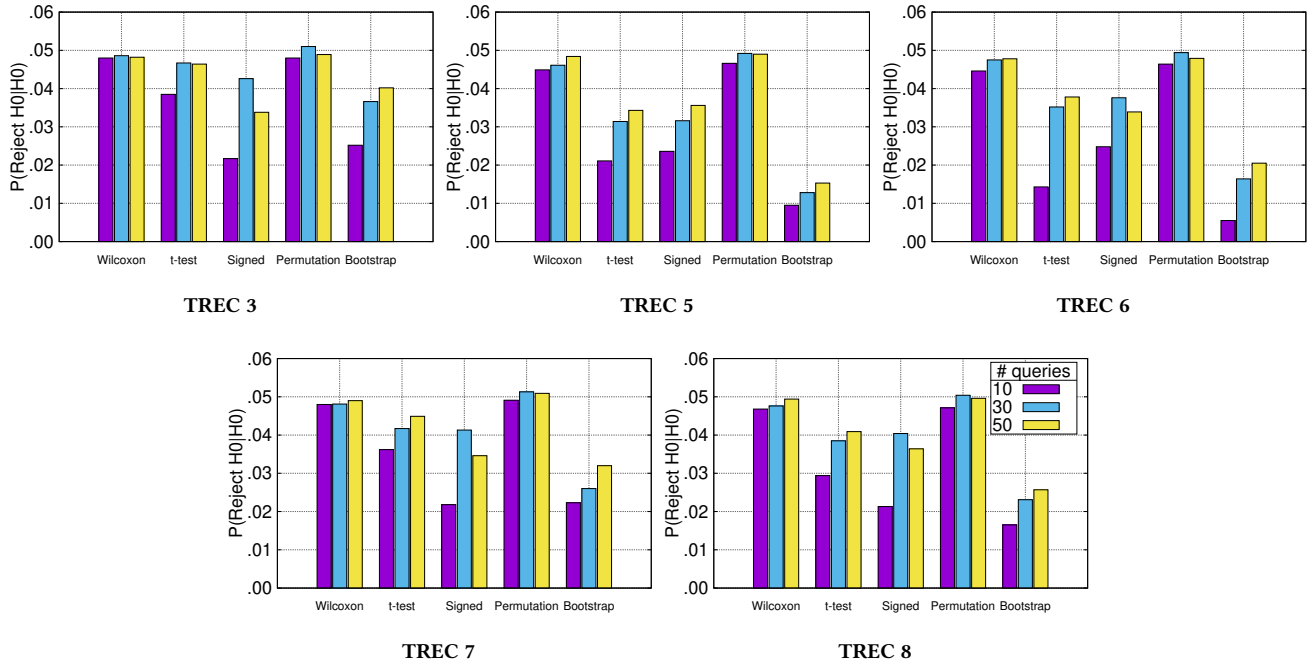

Figure 3: Average $P(Reject\ H0|H0)$ ($\alpha = 0.05$) in different TREC collections. For each collection, system and significance test, 10,000 experiments were run under H0, with different number of queries (increasing the sample size)

vast majority of the differences (in terms of AP) between pairs of TREC systems do not follow a Normal distribution according to Shapiro-Wilk normality tests (normality hypothesis is rejected with $\alpha = 0,05$ in most of the comparisons[7]). On the other hand, the permutation test assumes that observations are interchangeable under $H0$. For that to be true, equality of variances should hold. As argued in [2], this homogeneity requirement is well known but often overlooked and the permutation test should not be employed when the distributional requirements are not satisfied. On several applications of the permutation test, it has been observed that the robustness of this test is affected by the violation of this condition [2]. In Figure 6 (right) we show that the Brown-Forsythe test [4] (for equality of variances; $H0$: samples have homoscedasticity) is rejected with $\alpha = 0,05$ in the vast majority of the cases. More recently, Huang et al. [8] revisited this problem and showed that the permutation test for equality of means can be either too liberal or too conservative when samples have unequal variances. The heteroscedasticity of the paired samples also indirectly affects to the t-test: if the two groups do not have similar variances it is likely that the differences will not be normally distributed. In fact, Brown also found that the t-test is affected by the non homogeneity of variances, emerging the sign test as the most robust test in those situations [3].

Our experiments suggest that the permutation test is competitive when the benchmark has a low number of queries (10). However, the permutation test is substantially inferior to Wilcoxon when the benchmark has a higher number of queries (e.g. 50 queries, which is the typical situation in most IR experiments). The sign test is comparable to Wilcoxon in experiments with 30 or 50 queries but it is substantially inferior to Wilcoxon when the comparison is done with 10 queries. These experiments also show that the t-test and the Bootstrap test are the worst performers in terms of power. This superiority of the Wilcoxon test has been pointed out before in the literature. In [1], Blair and Higgins showed that the Wilcoxon Signed Rank test is more powerful that the t-test in most of the distributional situations, and increases its advantage with the sample size. Similar problems for the t-test on non-paired data were reported in [19].

By analysing the power of each test with increasingly larger query sets (10, 30 and 50), the reader can also observe how power increases with the sample size, which is a well-known result in Statistics.

4.3 Additional Experiments

We performed additional experiments where we tested other performance metrics –Normalized Discounted Cumulative Gain (NDCG) and P@10– and other variants of the significance tests (one-sided versions) under the same settings. These experiments confirmed the superiority of the Wilcoxon test and essentially revealed the same trends identified in the two-sided AP experiments reported above. Due to space constraints we only show a representative example. Figure 7 shows the power curves, computed using NDCG, that compare the (two-sided and one-sided) significance tests using one of the collections. The curves clearly demonstrate the ability of the Wilcoxon test to detect differences in NDCG between systems

[7] Shapiro-Wilk is considered to have good power [12].

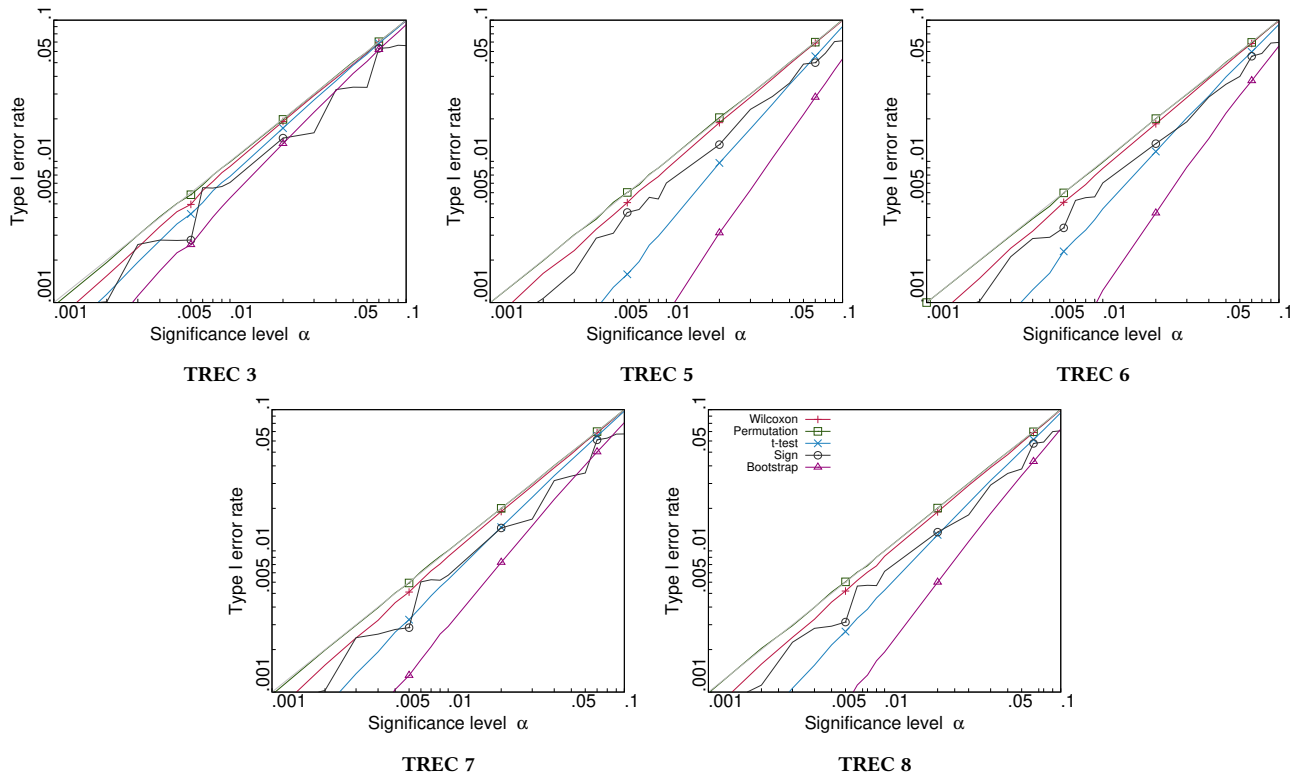

Figure 4: Type I error rates with varying significance levels in different TREC collections. For each collection, system and significance test, and significance level 10,000 experiments were run under H0, with 50 queries.

when the null hypothesis is false. Similar curves were obtained from the rest of the collections and for the one-sided variant of the tests.

Furthermore, we repeated the experiments but considering an alternative scenario that sometimes happens in IR evaluation. In some cases, system b is just a variation of an existing system a. For example, imagine that b aims at improving a's performance for certain types of queries (e.g., ambiguous queries or multi-topic queries). Under this setting, system a and b will yield the same performance for all queries but those that were manipulated by b. To further analyze the power of the tests under those situations (where the underlying model is only better for certain topics), we repeated the experiments by only improving a specific number of randomly selected query models (the rest of the models were not altered)[8]. The results for this experiment on TREC5 (2-sided for MAP) are shown in Figure 8. We found the same trends for all the datasets but space constraints preclude their inclusion here. The results corroborate the conclusions of previous experiments about the relative power of the tests. Additionally, this simulation shows that the power of the tests when improving 10 queries out of 50 is lower than when improving 10 queries out of 10 (see Figure 5).

[8]Remember that, as shown in Fig. 2b, the simulation is stochastic and, thus, a better query model does not always yields better effectiveness.

5 CONCLUSIONS

Our paper contributes with a novel approach to model the pattern of occurrence of relevant documents in search results rankings. Although the proposed models can support a number of IR tasks, we focused here on exemplifying the potential of these models to evaluate the use of tests of statistical significance in IR evaluation.

As shown by Sakai [15] and Carterette [5], the use of statistical tests in IR is pervasive, but IR practitioners do not have a clear method of choice. Our systematic comparison of significance tests for IR aims to fill this gap. The proposed method, which does not depend on the existence of scores, is free from the problems of previous comparisons [11, 20, 21, 24]. We worked with full control over $H0$ and modelled realistic and unbiased IR research conditions. Our conclusions agree with the ones of [11] and with established knowledge about the statistical tests [6].

We showed that Wilcoxon, and to a lesser degree, the sign test are the most powerful tests. IR practitioners should opt for the most powerful test. Choosing a low power test can lead to researchers to discard new innovations just because the test is not able to detect that the new model is indeed significantly better than the state-of-the-art. Of course, there is an important distinction between statistical significance and practical significance. No statistical test can tell you whether a given improvement is large enough to be put into production. In any case, choosing an inadequate test can

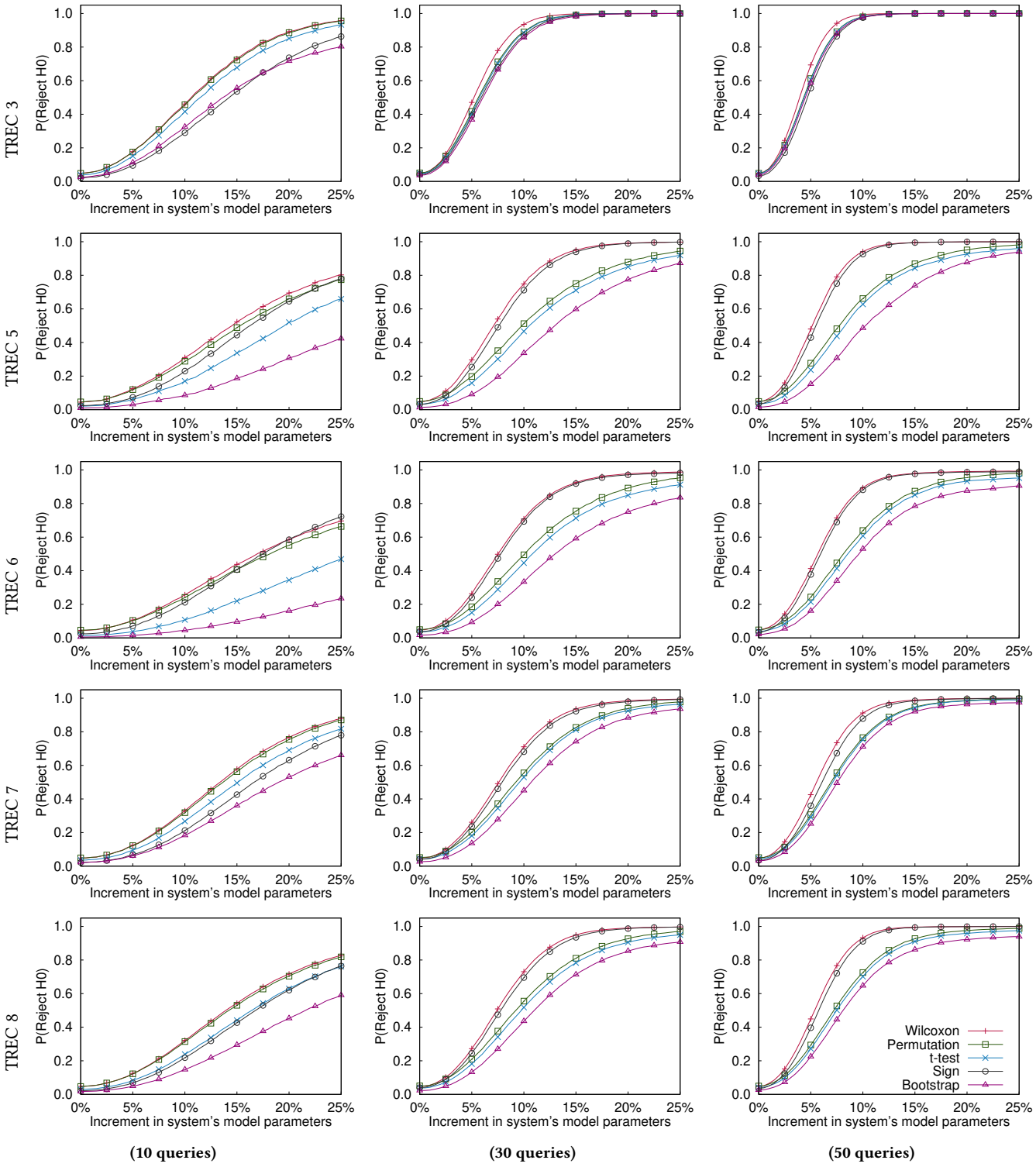

Figure 5: Average $P(Reject\ H0)$ ($\alpha = 0.05$) on different TREC collections. For each system and significance test, 10,000 experiments were performed and averaged. The experiments ranged from comparing equal systems (leftmost point, 0% increment) to comparing substantially different systems (25% increment) (in steps of 0.5%). Different columns correspond to different number of queries (increasing the sample size)

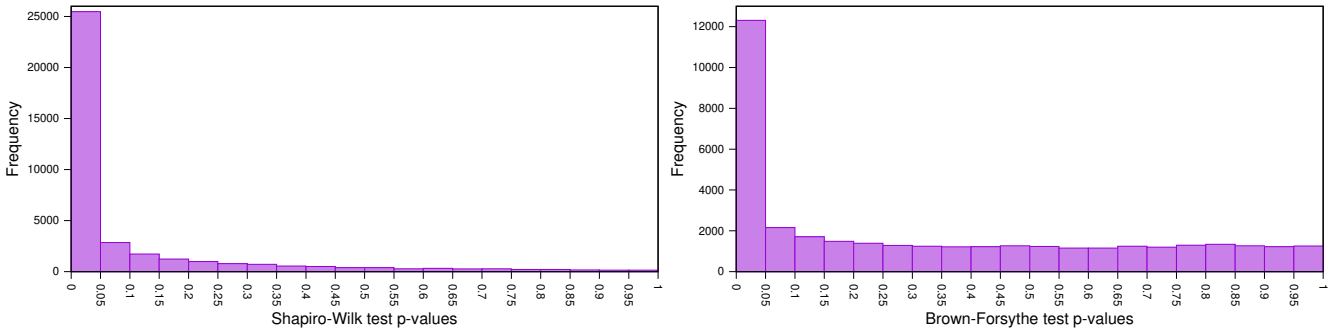

Figure 6: (left) p-values of Shapiro-Wilk normality test over AP differences between every pair of systems ($H0$ = values follow a Normal distribution); (right) Brown-Forsythe homoscedasticity test over AP between every pair of systems ($H0$ = values have the same variance).

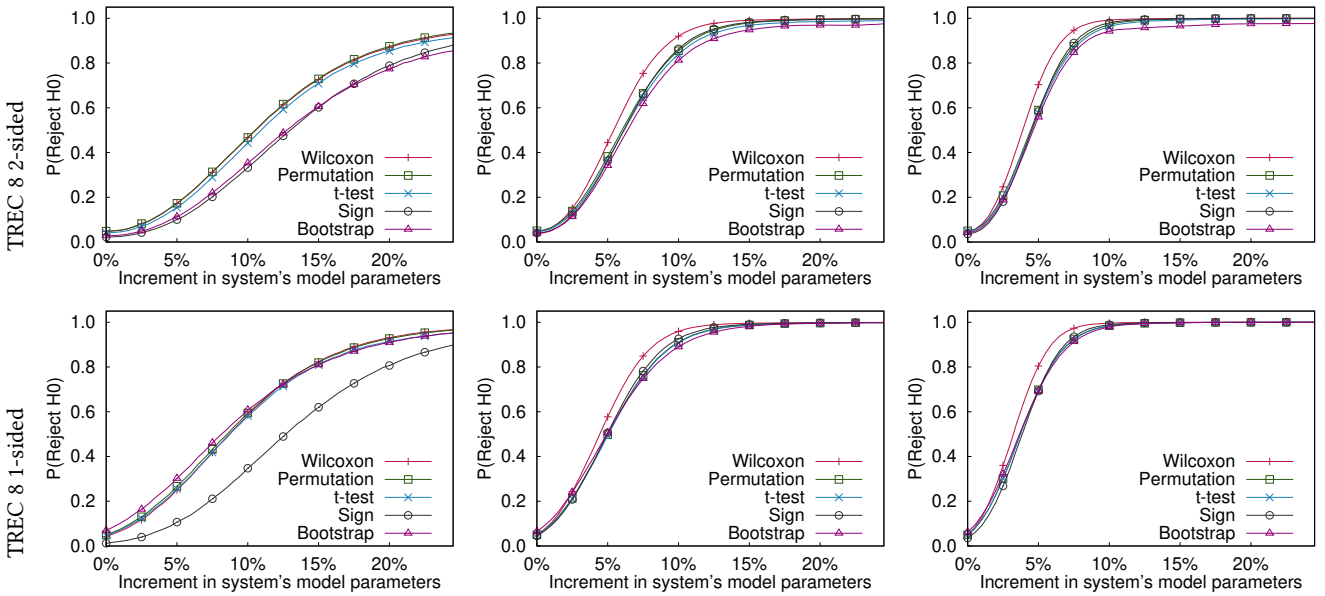

Figure 7: Average of the $P(Reject\ H_0)$ ($p < 0.05$) on TREC 8 for NDCG (averaged over 10,000 repetitions). The experiments ranged from comparing equal systems (leftmost point) to comparing substantially different systems (in steps of 0.5%).

slow the advance of IR research, where improvements are generally not dramatic.

As future work, we plan to extend the use of this methodology to other types of experimental comparisons. In particular, our simulation approach could be employed to support experiments comparing multiple systems, where a system is compared with many others. As argued by Sakai [17], other tests, such as the Tukey HSD test or the Bonferroni Correction, are required under this setting. Our models can naturally simulate the output of multiple systems and, thus, we can evaluate other tests that are adequate for comparing multiple systems and can support repeated measures.

ACKNOWLEDGMENTS

This work was supported by projects RTI2018-093336-B-C21, RTI-2018-093336-B-C22 (Ministerio de Ciencia e Innvovación & ERDF). The first and third authors thank the financial support supplied by the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G/01, ED431B 2019/03) and the European Regional Development Fund, which acknowledges the CITIC Research Center in ICT of the University of A Coruña as a Research Center of the Galician University System. The second author also thanks the financial support supplied by the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04, ED431C 2018/29) and the European Regional Development Fund, which acknowledges the

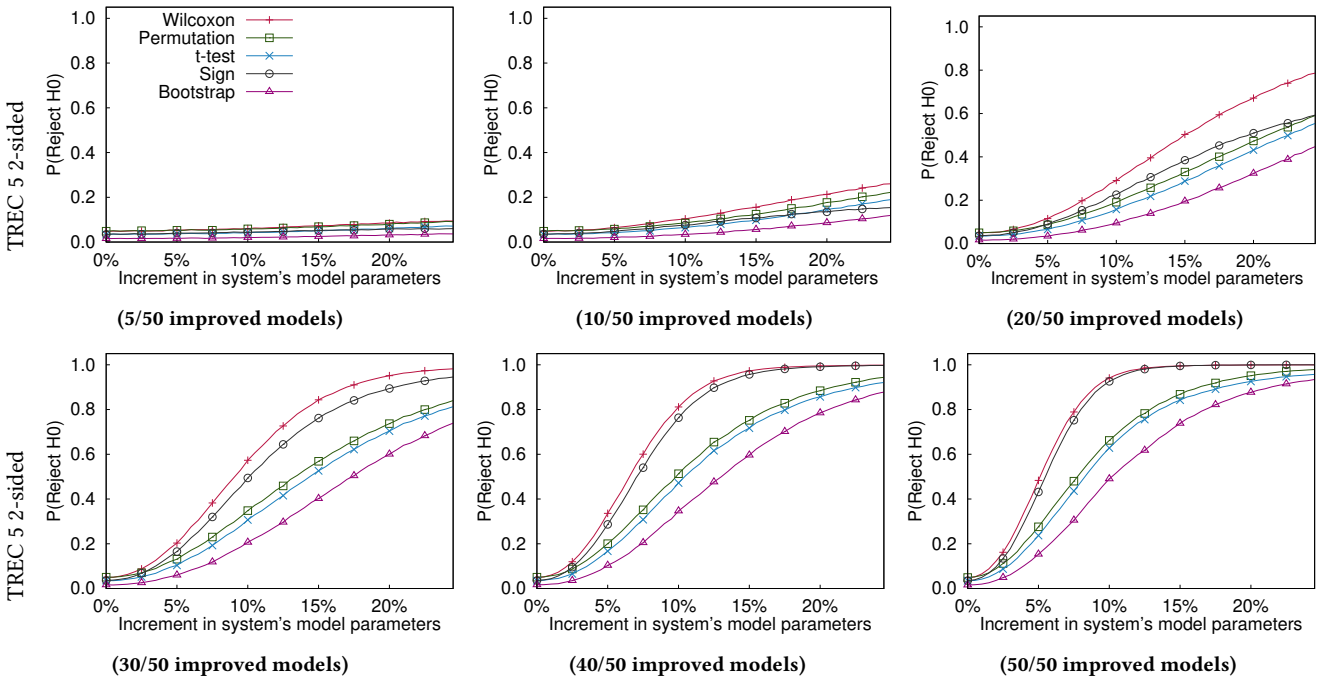

Figure 8: Average of the $P(Reject\ H_0)$ ($p < 0.05$) on TREC 5 for MAP (averaged over 10,000 repetitions) with different number of queries improved. The experiments ranged from comparing equal systems (leftmost point) to comparing substantially different systems by only improving (in steps of 0.5%) the specified number of query models

CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System.

REFERENCES

[1] R. Clifford Blair and James J. Higgins. 1985. Comparison of the Power of the Paired Samples t Test to That of Wilcoxon's Signed-Ranks Test Under Various Population Shapes. (1985). https://doi.org/10.1037/0033-2909.97.1.119

[2] Robert J. Boik. 1987. The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. 40, 1 (1987), 26–42.

[3] B. M. Brown. 1982. Robustness Against Inequality of Variances. (1982). https://doi.org/10.1111/j.1467-842X.1982.tb00834.x

[4] Morton B. Brown and Alan B. Forsythe. 1974. Robust tests for the equality of variances. (1974). https://doi.org/10.1080/01621459.1974.10482955

[5] Ben Carterette. 2017. Statistical Significance Testing in Information Retrieval: Theory and Practice. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017) *(SIGIR '17)*. ACM, 1387–1389.

[6] William Jay Conover. 1999. *Practical nonparametric statistics* (3rd ed.). Wiley.

[7] Gordon V. Cormack and Thomas R. Lynam. [n.d.]. Validity and Power of T-test for Comparing MAP and GMAP. In *Proceedings SIGIR '07* (2007). ACM, 753–754.

[8] Yifan Huang, Haiyan Xu, Violeta Calian, and Jason C. Hsu. 2006. To Permute or Not to Permute. 22, 18 (2006), 2244–2248. https://doi.org/10.1093/bioinformatics/btl383

[9] Evangelos Kanoulas, Keshi Dai, Virgil Pavlu, and Javed A. Aslam. 2010. Score Distribution Models: Assumptions, Intuition, and Robustness to Score Manipulation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2010) *(SIGIR '10)*. ACM, 242–249.

[10] Marco Marozzi. 2013. Nonparametric simultaneous tests for location and scale testing: A comparison of several methods. (2013). https://doi.org/10.1080/03610918.2012.665546

[11] Javier Parapar, David E. Losada, Manuel A. Presedo-Quindimil, and Alvaro Barreiro. 2020. Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology* 71, 1 (2020), 98–113. https://doi.org/10.1002/asi.24203

[12] Nornadiah Mohd Razali and Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. (2011). https://doi.org/doi:10.1515/bile-2015-0008

[13] Stephen Robertson, Evangelos Kanoulas, and Emine Yilmaz. 2013. Modelling Score Distributions Without Actual Scores. In *Proceedings ICTIR 2013* (2013). ACM, 20:85–20:92.

[14] Tetsuya Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (June 2014), 3–12. https://doi.org/10.1145/2641383.2641385

[15] Tetsuya Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2016) *(SIGIR '16)*. ACM, 5–14.

[16] Tetsuya Sakai. 2016. Two Sample T-tests for IR Evaluation: Student or Welch?. In *Proceedings SIGIR 2016* (2016). ACM, 1045–1048.

[17] Tetsuya Sakai. 2018. Multiple Comparison Procedures. In *Laboratory Experiments in Information Retrieval*, Springer (Ed.). Singapore, 59–80.

[18] Mark Sanderson and Justin Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings SIGIR 2005* (2005). ACM, 162–169.

[19] Shlomo S. Sawilowsky and R. Clifford Blair. 1992. A More Realistic Look at the Robustness and Type II Error Properties of the t Test to Departures From Population Normality. (1992). https://doi.org/10.1037/0033-2909.111.2.352

[20] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings CIKM 2007* (2007). ACM, 623–632.

[21] Julián Urbano, Harlley Lima, and Alan Hanjalic. 2019. Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019) *(SIGIR'19)*. ACM, 505–514.

[22] Julian Urbano and Thomas Nagler. 2018. Stochastic Simulation of Test Collections: Evaluation Scores. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018) *(SIGIR '18)*. ACM, 695–704.

[23] Ellen M. Voorhees. 2001. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2001) *(SIGIR '01)*. ACM, 74–82.

[24] Justin Zobel. 1998. How Reliable Are the Results of Large-scale Information Retrieval Experiments?. In *Proceedings SIGIR 1998* (1998). ACM, 307–314.