Inference at Scale

Significance Testing for Large Search and Recommendation Experiments

Ngozi Ihemelandu
Boise State University
Boise, Idaho, USA
ngoziihemelandu@u.boisestate.edu

Michael D. Ekstrand
Boise State University
Boise, Idaho, USA
ekstrand@acm.org

ABSTRACT

A number of information retrieval studies have been done to assess which statistical techniques are appropriate for comparing systems. However, these studies are focused on TREC-style experiments, which typically have fewer than 100 topics. There is no similar line of work for large search and recommendation experiments; such studies typically have thousands of topics or users and much sparser relevance judgements, so it is not clear if recommendations for analyzing traditional TREC experiments apply to these settings. In this paper, we empirically study the behavior of significance tests with large search and recommendation evaluation data. Our results show that the Wilcoxon and Sign tests show significantly higher Type-1 error rates for large sample sizes than the bootstrap, randomization and t-tests, which were more consistent with the expected error rate. While the statistical tests displayed differences in their power for smaller sample sizes, they showed no difference in their power for large sample sizes. We recommend the sign and Wilcoxon tests should not be used to analyze large scale evaluation results. Our result demonstrate that with Top-N recommendation and large search evaluation data, most tests would have a 100% chance of finding statistically significant results. Therefore, the effect size should be used to determine practical or scientific significance.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness**; **Retrieval efficiency**; *Relevance assessment*; **Recommender systems**.

KEYWORDS

evaluation, statistical inference

ACM Reference Format:
Ngozi Ihemelandu and Michael D. Ekstrand. 2023. Inference at Scale: Significance Testing for Large Search and Recommendation Experiments. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592004

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

1 INTRODUCTION

A key goal in information retrieval (IR) and related research is to make progress by promoting only those newly proposed algorithmic methods that truly improve the effectiveness of the state-of-the-art. Statistical significance testing plays an important role in achieving this goal. IR researchers use statistical significance tests to assess if an observed improvement is significant or literally due to random chance.

While there has been substantial discussion of the need for statistical significance in IR research [19], recent survey of papers published at ACM RecSys found that over half of the papers examined did not use any significance test to analyze their evaluation results [12]. One hurdle in addressing this is that there is little to no practical guidance on how to do inference for top-N recommendation experiments. There is, however, research and guidance on which statistical tests are the most appropriate to use for classic TREC-style experiments (small-scale ad-hoc search) [11, 16, 21, 22, 24, 29]. Our goal is to determine if that advice holds for top-N recommendation experiments, along with large search experiments that share many similar properties.

Top-N recommendation tasks and large-scale search tasks, such as MS-MARCO document or passage ranking [5, 13] have several distinguishing features: they have many topics (1, 000 or more vs. 50 for a TREC track); relevance data is sparse and incomplete compared with the (approximately) complete ground-truth relevance judgements obtained by pooling in TREC experiments; and in the top-N recommendation case, a few items are often relevant to many users resulting in a popularity skew, in contrast to a TREC experiment where documents are not concentrated to just a few topics.

In this paper, we empirically investigate how the pairwise significance tests — t-test, bootstrap, randomization, Wilcoxon signed-rank, and sign tests — which are commonly used to evaluate two IR systems using the same topic sets and which were studied in previous literature [16, 22, 24] behave with top-N recommendation system and large search evaluation data. We employ the simulation methodology proposed by Urbano and Nagler [25] to empirically compare these statistical tests under the full knowledge of the null hypothesis. Our study addresses the following questions:

- When two systems have equivalent performance (the null hypothesis is true), how frequently do the pairwise significance tests erroneously detect improvement in performance on large search and recommendation evaluation data?
- When one system outperforms another system (the null hypothesis is false), how powerful are the pairwise significance tests in detecting this improvement from large search and recommendation evaluation data?

We find that as the sample size increases, the false positive rates of sign and Wilcoxon tests increases. We also observe that as the sample size increases, the power of the tests increases, and the differences in the power amongst the tests diminish until there is no longer any distinguishable difference in their power. We find that Top-N recommendation and large search experiments have a 100% chance of finding statistically significant results even for very small effect sizes which may not be practically significant.

2 BACKGROUND AND RELATED WORK

The evaluation of information retrieval systems focuses on how well the system retrieves and ranks the retrieved documents. The Cranfield experiment is usually used for IR system evaluation [4, 26]. The effectiveness of a system is measured over a set of queries or topics using set of relevance judgements *qrel*. Let $E_1, E_2, E_3, \ldots, E_n$ be the effectiveness scores over n topics of the experimental method E and $B_1, B_2, B_3, \ldots, B_n$ be the effectiveness scores of the baseline method B. To compare their performance, we compute the difference of B_i and E_i such that $d_i = E_i - B_i$. We use the pairwise comparison statistical significance tests to determine if the mean of the observed difference $\bar{d}_i$ is significant or due to statistical random chance. The pairwise comparison statistical significance tests are commonly used when the data are paired, as when we evaluate two IR systems using the same topic sets [20].

Five pairwise tests and their null hypotheses are:

Student's Paired t-test The mean difference is zero [7].
Bootstrap shift test E and B are samples from the same distribution [14].
Randomization test System E and system B have identically-distributed effectiveness scores [2].
Sign test The median(d)=0 [27].
Wilcoxon signed rank test d symmetric with median 0 [27].

The sign and Wilcoxon tests are used for IR experiments because they were recommended in the early account of statistical significance testing in IR [18, 29]. Considering the mean or the median is inconsequential unless the data is skewed substantially [17].

A number of studies have investigated the suitability of these statistical tests for the analysis of IR evaluation data particularly for the results of TREC-style experiments [11, 16, 21, 22, 24, 29]. Smucker et al. [22] used the randomization test as ground truth for comparing pairs of TREC runs and compared the other tests to it. They recommended the discontinuation of the Sign and Wilcoxon tests because they tend to disagree with the randomization test while the bootstrap and t-test agreed with it. Urbano et al. [24] and Parapar et al. [16] trained simulations to mimic evaluation results from historic TREC evaluations. Simulation allowed them to control whether the null hypothesis was true or not, and compute actual error rates and power. Parapar et al. [16]'s method simulates new runs for the same topic while Urbano et al. [24]'s approach simulates new topics for the same system. Based on their findings, Urbano et al. [24] recommended the use of the t-test while Parapar et al. [16] recommended the use of the sign and Wilcoxon tests. There is some debate about these conflicting results [15], whether they arise from Urbano et al.'s use of parametric distributions, but Urbano et al. [23] argue that their methods would favor the Wilcoxon test if it was appropriate.

We treat the topic evaluation scores of a system as observations of a random variable following Sakai [20]. Therefore, we adopt the method proposed by Urbano and Nagler [25] for simulation.

3 DATA AND METHODS

We used evaluation results from multiple tasks (passage ranking and top-N recommendation) to study statistical test behavior. Each evaluation results dataset consists of a set of systems, a set of evaluation requests (a topic for search, and a user for recommendation), and effectiveness scores. We selected these evaluation datasets because the typify the datasets used in large search and top-N recommendation experiments; the size of the evaluation requests ranges from 943 to 32, 509 (substantially more than the 50 evaluation requests typically used for the ad-hoc search TREC track), and the relevance data is sparse and highly incomplete.

Table 1 summarizes our evaluation results dataset.

Table 1: Summary of the evaluation results studied.

Dataset	Task	Systems	Requests
ML-100K	Recommendation	58	943
ML-25M	Recommendation	58	32,509
AZ-Video	Recommendation	58	6,000
MS MARCO	Passage ranking	63	6,980

3.1 Large-Scale Search

We used a collection of MS-MARCO dev set runs provided by the MS-MARCO team [1]. We evaluated the runs with the official evaluation script they provided to participants[1]. MS-MARCO runs use the reciprocal rank (RR) metric for effectiveness, as reported by the official evaluation script they provided to participants.

3.2 Top-N Recommendation System

We generated evaluation scores for top-N recommendation by training four recommender system algorithms with different hyperparameters values (see Table 3) using LensKit for Python (version 0.13) [6]. The combinations of algorithms and associated hyperparameters values gave a total of 58 systems.

We trained the algorithms on three publicly-available data sets (two from MovieLens [9] — ML-100K and ML-25M — and AZ-Video from Amazon [10]). Table 2 summarizes these datasets.

Table 2: Summary of data sets.

Dataset	# ratings	# requests	Density
ML-100K	100,000	1,000	6.3%
ML-25M	25,000,000	162,000	0.26%
AZ-Video	583,933	29,756	0.01%

For consistency with MS-MARCO, we used the RR metric to assess the quality of the recommendations generated for each user by the different trained models. We also computed nDCG scores

[1]https://github.com/microsoft/MSMARCO-Passage-Ranking/blob/master/ms_marco_eval.py

but omit them for reasons of space. The evaluation of the generated recommendations resulted in a set of "runs" where a run represents a trained model and comprise of the evaluation scores for a set of users.

To train and evaluate an algorithm for a dataset, we split the dataset by partitioning the set of users into 5 sets of test users. For each set of test users, we selected 20% of a test user's interactions for testing and used the rest for training, along with data from the other users. For each algorithm, we:

- trained on the training data.
- generated 100 recommendations for each test user.
- used the RR metrics to evaluate the generated recommendation lists.

The selection of 20% of a test user's rows implied that users with fewer than 5 ratings were not included in the test data.

Table 3: Recommender algorithms used and their hyperparameter settings.

Algorithm	Hyperparameter	Values
Item k-NN	nnbrs	5, 10, 20, 30, 50
User-based k-NN	nnbrs	5, 10, 20, 30, 50
ALS BiasedMF	als-features	5, 10, 25, 50, 100, 200, 300, 500
	als-iterations	5, 10, 20
ALS ImplicitMF	als-features	5, 10, 25, 50, 100, 200, 300, 500
	als-iterations	5, 10, 20

3.3 Experiments

We designed the experiment to compare the false positive rates and power of the two-tailed statistical tests on the four datasets for sample sizes $n = 25, 50, 100$ to be consistent with sample sizes used in [16, 24] and $n = 500, 1000, 5000, 10000, 20000$ to get closer to the sample sizes used in large search and recommendation experiments. We used the methodology proposed by Urbano and Nagler [25], a generative stochastic simulation model which consists of two parts: (1) a system's effectiveness score distribution (marginal distribution for the system) and (2) a bivariate copula that models the dependence between pairs of runs (joint distribution). We simulated new effectiveness scores from the fitted model over arbitrarily topics.

To fit the model:

(1) Fit a marginal distribution F_B to each run B in the evaluation dataset such that $B \sim F_B$. We would refer to these as baseline runs. These marginal distributions are parametric and non-parametric which take the form of Truncated Normal, Beta, Beta-Binomial and discrete kernel smoothing distributions.
(2) Select a pair of runs ($B \sim F_B$, $E \sim F_E$) whose difference in mean scores are closest to an effect size δ and modify F_E such that it is transformed with a new mean $\mu_E = \mu_B + \delta$.
(3) Use the CDF of a baseline run F_B to transform B to pseudo-observations U_B such that $U_B \sim \text{Uniform}(0, 1)$.
(4) Fit the copula C to every pair of U_B.

To measure the false positive (type I) error rate we simulated pairs of runs from the fitted model (in this scenario the null hypothesis is true) as follows:

(1) Draw new pseudo-observations (V_B, V_E) from the fitted copula.
(2) Apply the inverse CDF of the marginal B to V_B, V_E to get the final effectiveness scores $B = F_B^{-1}(V_B)$ and $E = F_B^{-1}(V_E)$. Using the same marginal ensures that the effectiveness is the same between systems.

We simulated n new runs and computed the 2-tailed p-values for each of the statistical tests, repeating this process 10,000 times for each combination of dataset, metric, and sample size for a total of 180,000 trials. A statistically significant result by any of the test is counted as false positive against the test.

To measure the power, we generated runs (B, E) with different effect sizes δ such that $\mu_E = \mu_B + \delta$ (in this scenario the null hypothesis is false). We generated the runs as follows:

(1) Select a fitted copula with baseline runs that have a difference in means close to δ. The copula is associated with a baseline run B and a transformed run E.
(2) Draw pseudo-observations (V_B, V_E) from the selected copula.
(3) Apply the inverse CDF of B to V_B, and the inverse CDF of E to V_E to get effective scores for new requests ($B = F_B^{-1}(V_B)$, $E = F_E^{-1}(V_E)$).

We simulated pairs of effectiveness scores of size n with effect sizes δ and computed the 2-tailed p-values for each of the statistical tests. We repeated this process 2,500 times for every combination of δ (in the range $[0.01, 0.02, \ldots, 0.1]$), sample size n, metric, and dataset which gave us a total of 1,400,000 trials. Under this model, the null hypothesis is false. Therefore, any statistically significant result by any of the test will count as true positive (power).

4 FINDINGS

Figures 1 and 2 show the false positive error rates of the tests for different sample sizes at significance levels $\alpha = 0.01, 0.05, 0.1$ and the p-values calibration of the tests for a broader range of α respectively. When the null hypothesis is true (two systems have equivalent performance), a significance test is expected to be well calibrated (close to the diagonal) with false positive error rate that is equal to the significance level α.

From figure 1, we observe the following behavior of the significance tests on the different datasets:

- For small sample sizes, the sign test makes less error than expected, the Wilcoxon test is close to the expected behavior, and the bootstrap makes more error than expected.
- As the sample size increases, the bootstrap approaches the expected behavior while the sign and Wilcoxon test starts making more errors than expected. We observed a much more higher than expected error rates with the nDCG metric for the sign and Wilcoxon tests but omit them for reasons of space.
- The t-test and randomization tests are close to the expected behavior for all sample sizes.

Further, figure 2 shows that at $n = 50$ the sign test has lower than expected false positives across a range of significance thresholds, while at $n = 20000$, both sign and Wilcoxon have higher than expected error.

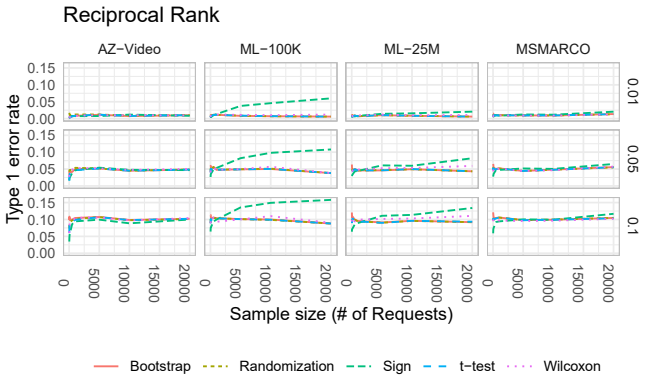

Figure 1: False positive rates of 2-tailed tests. Expected behavior: The false positive rate of a test should be directly proportional to a specified significance level ($\alpha = 0.01, 0.05, 0.1$).

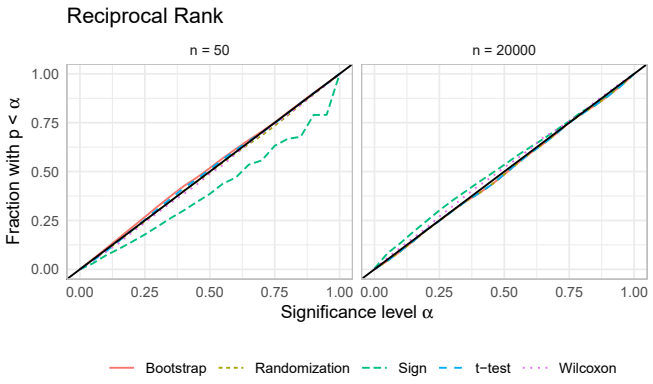

Figure 2: The calibration of the p-values of the statistical significance tests. A test is well calibrated if it is close to the diagonal. Fraction of mean difference with a p-value below α is approximately the same as α.

Figure 3 shows the power for the two-tailed tests at significance level $\alpha = 0.05$. We observe the following behavior of the significance tests on the different datasets:

- For small sample size (n=50), there are differences in the power of the tests. The sign and Wilcoxon tests are the most powerful and the t-test the least powerful for the RR metric.
- As the sample size increases the power of the tests increases, and the observed differences in their power decreases such that there is no distinguishable difference.

With large sample sizes, all tests readily find extremely small effects.

5 DISCUSSION AND CONCLUSION

Our empirical study of the false positive rates of the pairwise statistical tests with Top-N recommendation and large search evaluation data, showed that as the sample size increases, the false positive rates of the sign and Wilcoxon tests increases. Under the null hypothesis, the difference distribution is expected to be symmetric

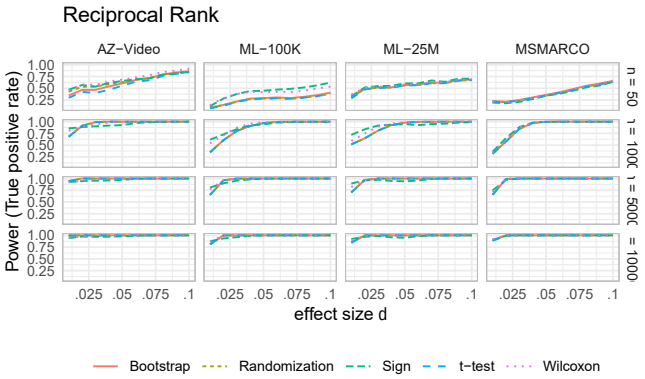

Figure 3: Power of statistical tests as a function of effect size and sample size at significance level: 0.05.

about zero. Slight deviation from symmetry about zero is usually inconsequential and not statistically significant for small sample sizes. However, we observe that when the sample size is large, the sign and Wilcoxon tests find these slight deviations from perfect symmetry about zero in the difference distribution as statistically significant. This result is in agreement with findings from Smucker et al. [22] and Urbano et al. [24] that the Wilcoxon and sign tests are unreliable for the analysis of mean effectiveness. Our results demonstrates that this unreliability has even more serious consequences with the sample sizes found in large search and recommendation experiments. Therefore, we recommend that the sign and Wilcoxon tests should not be used for analyzing recommendation and large search evaluation data.

This study also analyzed the power of the pairwise significance tests in detecting performance improvement from large search and recommendation evaluation data. We found that for small sample size the power of the pairwise significance tests are distinguishable but as the sample size increases, the difference in the power amongst the tests begin to diminish until there is no longer any distinguishable difference in their power (see 3). While previous research [16, 24] made recommendations that were based on findings that some tests were more powerful than others, our result demonstrate that with Top-N recommendation and large search evaluation data, most tests would have a 100% chance of finding statistically significant results. Therefore, we can be confident that the effect size is precise and not actually 0. However, we need to decide if the effect size is scientifically or practically meaningful. We recommend that both the p-value, and effect size be reported [8, 28], and used for decision making.

Further research is needed that provide statistical techniques that provide meaningful results in light of the specific problems experienced in top-N recommendation and large search evaluation.

ACKNOWLEDGMENTS

This work partially supported by the National Science Foundation under Grant IIS 17-51278. Computation performed on Borah [3]. MS-MARCO runs generously provided by Bhaskar Mitra and Nick Craswell.

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[2] George EP Box, William H Hunter, Stuart Hunter, et al. 1978. *Statistics for experimenters*. Vol. 664. John Wiley and sons New York.

[3] Kelly A Byrne. 2020. Borah: Dell HPC Intel (High Performance Computing Cluster). (2020).

[4] Cyril Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*. MCB UP Ltd.

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1566–1576.

[6] Michael D Ekstrand. 2020. Lenskit for python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2999–3006.

[7] Ronald Aylmer Fisher. 1936. Design of experiments. *British Medical Journal* 1, 3923 (1936), 554.

[8] Norbert Fuhr. 2018. Some common mistakes in IR evaluation, and how they can be avoided. In *Acm sigir forum*, Vol. 51. ACM New York, NY, USA, 32–41.

[9] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[10] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[11] David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 329–338.

[12] Ngozi Ihemelandu and Michael D Ekstrand. 2021. Statistical Inference: The Missing Piece of RecSys Experiment Reliability Discourse. *arXiv preprint arXiv:2109.06424* (2021).

[13] Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. 2021. Significant improvements over the state of the art? a case study of the ms marco document ranking leaderboard. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2283–2287.

[14] Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

[15] 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 *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 655–664.

[16] 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.

[17] Kandethody M Ramachandran and Chris P Tsokos. 2020. *Mathematical statistics with applications in R*. Academic Press.

[18] V Rijsbergen and C Keith Joost. 1979. Information Retrieval Butterworths London. (1979).

[19] 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*. 5–14.

[20] Tetsuya Sakai. 2018. Laboratory experiments in information retrieval. *The information retrieval series* 40 (2018).

[21] Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 162–169.

[22] Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 623–632.

[23] Julián Urbano, Matteo Corsi, and Alan Hanjalic. 2021. How do metric score distributions affect the type I error rate of statistical significance tests in information retrieval?. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 245–250.

[24] 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*. 505–514.

[25] Julián 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*. 695–704.

[26] Ellen M Voorhees. 2001. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*. Springer, 355–370.

[27] Dennis Wackerly, William Mendenhall, and Richard L Scheaffer. 2014. *Mathematical statistics with applications*. Cengage Learning.

[28] Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. 2019. Moving to a world beyond "p< 0.05". , 19 pages.

[29] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 307–314.