Statistical Inference: The Missing Piece of RecSys Experiment Reliability Discourse

Ngozi Ihemelandu[1], Michael D. Ekstrand[2]

[1]*People and Information Research Team, Boise State University, 1910 University Drive, Boise, Idaho, USA, 83725-2055*

[2]*People and Information Research Team, Boise State University, 1910 University Drive, Boise, Idaho, USA, 83725-2055*

Abstract

This paper calls attention to the missing component of the recommender system evaluation process: Statistical Inference. There is active research in several components of the recommender system evaluation process: selecting baselines, standardizing benchmarks, and target item sampling. However, there has not yet been significant work on the role and use of statistical inference for analyzing recommender system evaluation results.

In this paper, we argue that the use of statistical inference is a key component of the evaluation process that has not been given sufficient attention. We support this argument with systematic review of recent RecSys papers to understand how statistical inference is currently being used, along with a brief survey of studies that have been done on the use of statistical inference in the information retrieval community. We present several challenges that exist for inference in recommendation experiment which buttresses the need for empirical studies to aid with appropriately selecting and applying statistical inference techniques.

Keywords
LaTeX class, Evaluation, statistical inference, significance tests, significant results

1. Introduction

It is widely recognized that the use of appropriate statistical inference techniques should be used to analyze, interpret, and report the results of evaluations and experiments, including evaluations of recommender systems [1]. These techniques come in many forms, including point estimation, interval estimation, and hypothesis testing, but analysis needs to go beyond merely computing metrics to determine if observed metrics represent genuine effects. [1] In this paper we consider the state of statistical inference in recommender systems evaluation, arguing that identifying and documenting best practices for statistical analysis is a vital and oft-overlooked component of the discussion on how to improve the rigor, reproducibility, and reliability of recommender systems evaluation results.

[1]The video accompanying this paper is available at https://piret.info/pubs/2021/Perspectives21-StatisticalInference

We focus primarily on statistical inference for one of the most common goals of recommender systems research: to demonstrate an improvement in effectiveness over the current state of the art. This could be by developing a new recommendation technique that is more effective at some recommendation tasks than previously-known techniques, or by modifying an existing approach. To assess if the measured improvement of the new method over the state-of-the-art is substantial and not just a result of random chance, we typically use a hypothesis test (null hypothesis significance testing, or NHST) for the null hypothesis that there is no difference between the two methods' effectiveness; sometimes confidence intervals or Bayesian inference techniques may be employed instead of or in addition to an NHST.

There has been significant research on evaluation strategies for this research goal. Dacrema et al. [2] showed in their systematic analysis of deep learning approaches for top-N recommendation tasks that many claims of improved performance over a baseline may be illusory. There are many design points in a recommender experiment that can affect its rigor and reliability; Dacrema et al. focused specifically on the choice and tuning of baselines in the evaluation process. They found that many measured improvements disappear when the baseline algorithms are properly tuned: that is, better choice of hyperparameters and model options can cause the baseline to perform just as well as the proposed new method.

Other authors have considered the effects and sought to develop best practices for other design choices in an evaluation. Rendle et al. [3] argue for standardized benchmarks, by which they mean datasets with well-defined train–test splits and evaluation protocols for specific tasks (e.g. prediction). They state that although well-defined benchmarks exist for comparing prediction algorithms, there are not standardized benchmarks for other tasks such as ranking. They argue that empirical findings reported in research papers are questionable unless they were obtained on standardized benchmarks where — as recommended by Dacrema et al. — baselines have been tuned extensively. Cañamares and Castells [4] bring attention to an offline evaluation setup component — target item sampling — that is not always explicit and has received little attention in the quest for seeking an evaluation procedure. They show that different target subsets can lead to different evaluation outcomes. Sun et al. [5] work shed light on the issues – unreproducible evaluation and unfair comparison – which they attribute to the unavailability of effective benchmarks for evaluation. They investigated the evaluation rigorousness (reproducibility and fairness) in recommendation by analyzing the influence of different factors on recommendation performance through a holistic empirical study. The result of their study corroborates the findings of Dacrema et al. [2].

However, there has not yet been much attention to appropriately selecting and applying statistical inference techniques to the metrics that result from these evaluations. Shani and Gunawardana [6] discuss general ways of performing significance testing using widely-known statistical methods, but to our knowledge there have not yet been empirical studies on the use of statistical inference for analyzing evaluation results, as there has been for TREC-style search experiments (see Section 3). Evidence-based guidance on best practices for analyzing and reporting results is therefore lacking. The current use, or lack thereof, of various techniques for recommender system experimental results is also an open question.

Our central claim in this paper is that the RecSys community does not currently pay sufficient attention to the choice and use of statistical techniques, and discussions such as the one at this workshop needs to consider the role of inference and develop best practices for rigorous

analysis of evaluation results. We support this argument with a systematic review of recent RecSys papers to understand how statistical inference is currently being used, along with a brief survey of studies that have been done on the use of statistical inference in the information retrieval community (particularly for analyzing TREC search effectiveness metrics). We identify several challenges that exist for inference in recommendation experiments, and call on the community to attend to this issue and work with us to fill this important gap in the literature on reliable evaluation of recommender systems.

2. Systematic Review of Statistical Inference in RecSys

We begin by assessing current practices in statistical inference for recommender system evaluations. Our study is inspired by that of Sakai [7], who conducted a systematic review of 840 SIGIR full papers and 215 TOIS papers published between 2006 and 2015. Their goal was to identify what types of statistical test IR researchers use, how they report or fail to report on significance test results, and how the reporting practices may have changed over the last decade.

They found that of the 862 papers selected for the survey about 28-30% do not report significance test results; for the comparison of two IR systems, 61-66% of these papers use the paired t-test; 20-23% use the Wilcoxon signed rank test; 4-5% use the randomisation test; 3-4% use the sign test; and 1% use the bootstrap test. They also found that the paired t-test was more common in recent years while Wilcoxon test decreased in popularity.

To get a first look at current RecSys statistical practices, we conducted a systematic review only for long and short RecSys papers that proposed new or enhanced algorithmic methods and compared their performance to that of baselines (state-of-art). Hence, proposed new methods that were not compared to baselines were not selected. Our survey is limited to papers published in 2019 and 2020.

2.1. Survey Methods

The main focus of this systematic survey is to examine how statistical significance tests are used by researchers working on papers proposing new or enhanced recommender algorithms.

We selected full and short papers from RecSys 2019–2020 that meet the following criteria:

- The paper proposed a new or enhanced algorithmic method for some recommendation task.
- The effectiveness scores for the baselines and new/enhanced method were recorded.

We coded the selected papers as specified below (The coding was done in the listed order. That is, if the paper does not meet the first criteria, the second criteria is checked etc.):

Used specified test The paper mentioned the name of the test used along with the significance level (α) or p-value. We also recorded which test it used.

Used confidence interval The paper reported confidence intervals or indicated the standard error for the estimated metric scores for the new method as well as the baseline.
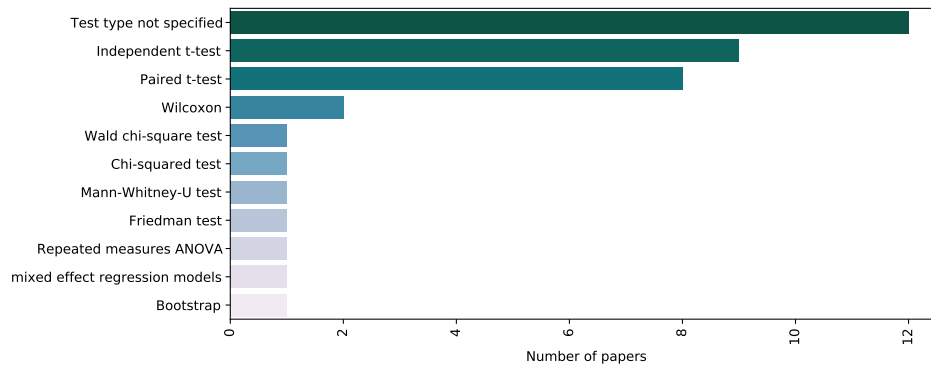
Table 1

Long and short algorithm papers by year

Year	# long papers	# short papers	long + short
2020	25	21	46
2019	36	29	65
Total			111

Table 2

Paper count and percentage for each category

Category	# Papers	% of Papers
Used specified test	26	23
Used confidence interval	8	7
Used unspecified test	12	11
No significance test	65	59

Figure 1: Distribution of selected papers coded as – used significance test.

Used unspecified test The paper did not specify which test was used but claimed statistical significance or specified p-value $<$ significance level (α) or the calculated test statistics.

No significance test The paper did not seem to test the results for significance.

2.2. Results and Discussion

Out of the 146 RecSys long and short papers examined, we found 111 papers that proposed new or enhanced recommender algorithms for which we expect significance testing to be used to analyze the evaluation result. See Table 1 for the break down of the selected papers by year.

Table 2 shows the classification of the 111 selected papers, and Fig. 1 shows the distribution by the test type of the set of selected papers labeled as "used significance tests". We found that over half of the papers proposing a new algorithmic method did not seem to use any significance test to analyze their evaluation results; a substantial portion of those who claim significance did not specify a test.

These results show that there is currently a lack of rigorous statistical analysis and reporting in the evaluations published in RecSys. While we do not have an explanation as to why there is this gap, we believe it needs to be filled if we are to go from observed differences in metrics to reliable knowledge.

3. Statistical Inference in Information Retrieval

In addition to Sakai's study of existing practice, several studies in the information retrieval (IR) community have addressed the use of statistical inference for system comparison experiment, attempting to identify which statistical techniques are appropriate to use for the analysis of the evaluation results in IR systems comparison, particularly for the results of TREC-style experiments.

Smucker et al. [8] used results from historical TREC runs to study the agreement between different pairwise significance tests. Using root mean squared error (RMSE) to compare the p-values produced by five different tests, they found that the randomization, bootstrap, and t-tests all agreed with each other (producing very similar p-values) while the Wilcoxon and sign tests neither agreed with the other tests nor each other. They then used the randomization test as ground truth to estimate the false positive and false negative rates of the Wilcoxon and Sign tests, finding that both tests have high false positive and false negative rates when the difference in system effectiveness (evaluation metric) is small. They recommend that researchers wanting a distribution-free test should use the randomization test with the test statistic of their choice, and recommended discontinuing use of the Wilcoxon or sign tests for IR evaluation data analysis.

Urbano et al. [9] and Parapar et al. [10] used simulations to produce per-topic evaluation scores rather than directly using the recorded metrics. They fit generative probabilistic models to the metric distributions from historic TREC runs (to ensure realism) and sampled from these models, allowing them to directly control the actual difference (or lack thereof) between systems and measure the error rates of different statistical tests. One of the key differences in their approaches is the simulation architecture: Parapar et al. [10] simulated the utility of individual retrieved documents, while Urbano et al. [9] modeled the joint distribution between pairs of effectiveness scores. Both simulation designs enabled them to directly assess the accuracy of the p-values produced by the various significance tests, and to measure their false positive rates and statistical power.

Urbano et al. [9] found that the Wilcoxon and sign tests have more false positives than expected, especially at low significance levels, and that this error is more pronounced as the sample size increases. The bootstrap test exhibits similar behavior (making more false positive rates than expected) with small sample sizes but starts behaving as expected as the sample size increases. The randomization test behaves better than the bootstrap, Wilcoxon and sign tests and approaches the expected behavior as the sample size increases. The t-test behaves as expected even for small sample size. They also found that for large sample sizes the randomization, bootstrap and t-test all agree, concurring with the results of Smucker et al. [8].

They also found that the sign test is consistently less powerful than other tests while the bootstrap test is usually the most powerful, especially with small samples. With large sample

sizes, all tests except the sign tests exhibited nearly-identical power. Since the t-test was well-behaved as in terms of both the false positive rate (even for small samples) and power, Urbano et al. recommend its use as the best choice for mean effectiveness in IR evaluations, and the randomization test for test statistics other than the mean. Like Smucker et al., they discourage use of the Wilcoxon and sign tests for IR evaluation results.

Parapar et al. [10] came to different conclusions than Urbano et al.. Their simulations showed that the Wilcoxon and randomization tests have the expected false-positive rate behavior while the t-test, the sign test, and the bootstrap did not behave as expected. They also found that the sign test and Wilcoxon test have more statistical power than the other tests. Therefore they recommend the use of the sign test and Wilcoxon test for the analysis of IR evaluation results.

All three papers had the goal of producing recommendations for appropriate significance tests to apply when comparing IR systems. Smucker et al. [8] and Urbano et al. [9] made recommendations that were similar, while Parapar et al. [10] arrived at a completely different recommendation.

Both Urbano et al. [11] and Parapar et al. [12] have followed up and attempted to understand this discrepancy in their conclusions, but there is not yet clarity on which is the more reliable recommendation.

4. Gaps for Fixing RecSys Evaluation Practice

Whichever evidence produces the more reliable recommendation for IR evaluation settings studied in the previous section, it may not be feasible to just apply that recommendation to RecSys evaluation. There are some key differences between TREC ad-hoc retrieval evaluation and the recommender system evaluation. Some of these key differences — which do not only affect recommendation, as many are shared with actual deployments of search engines outside the TREC context — include:

- The sample size of the test collection in a traditional TREC Cranfield experiment is quite small — often 50 topics, particularly in the data sets studied — while the typical sample size of a RecSys evaluation is $> 1,000$ users.
- In typical RecSys evaluation data, a few items are known to be relevant to many users, resulting in a long-tailed distribution of user ratings over items. This is in contrast with TREC evaluation where documents are not concentrated to just a few queries.
- In TREC evaluation, the ground-truth relevance judgement which are assumed to be (approximately) complete, while the user feedback used in RecSys evaluations form a sparse and highly incomplete picture of item-user relevance.

We want to call particular attention to sample size, as it is a key factor that impacts the statistical power of a significance test (the ability of the test to detect significance in the presence of a real effect). The statistical power of a significance test increases as the sample size increases; therefore, by increasing the sample size, any measured improvement can be found to be significant by any significance test even when the size of the measured improvement is so small that it is not operationally meaningful.

Statistical biases are another factor that may influence the outcome of significance test for RecSys evaluation data. It has become well known that biases such as sparsity and popularity biases in RecSys evaluation data considerably distort the evaluation measures [13, 14, 15, 16, 4]. Bellogín et al. [17] showed that the long-tailed distribution of RecSys evaluation data has a drastic effect on how recommendation algorithms compare to each other. The hypothesis test does not account for these biases hence, this distortion can ultimately influence its outcome. It isn't clear whether this should be fixed as a part of inference, or as a corrective stage before or after inference, but it remains a gap in the ability to accurately evaluate system performance that needs to be addressed.

There are also on-going discussions on the inadequacies of statistical significance testing. McShane et al. [18] states that the widespread crisis in the biomedical and social sciences with published findings failing to replicate at an alarming rate maybe associated with claims of huge effects from tiny interventions, citing $p < 0.05$ as the primary evidence. A group of 72 researchers representing a wide range of disciplines (psychology, economics, sociology, anthropology, medicine, epidemiology, ecology, and philosophy) and statistical perspectives have proposed a change in the p-value threshold for a "statistically significant" result from 0.05 to 0.005 for claims of discoveries of novel effects [19]. They recommend that results currently called "statistically significant" that do not meet the new threshold would be called suggestive and treated as ambiguous as to whether there is an effect. However, McShane et al. [18] state that this proposal is insufficient to overcome the current crisis with the inability to replicate experiment results. They recommend abandoning the null hypothesis significance testing paradigm entirely and just use p-values as one of many pieces of information to cite as evidence for a novel effect claim.

Translating this discussion back to information retrieval, Sakai [20] recognizes that statistical significance testing is not enough and provides suggestions on how IR researchers should report effect sizes and confidence intervals along with p-values, in the context of comparing IR systems using test collections. Carterette [21] advocates for the use of the t-test even though their analysis showed that a p-value cannot have any objective meaning. They believe it is still useful for many of the purposes they are currently used for. They however, recommend that in the long term, IR experimental analysis should transition to a fully Bayesian modeling approach.

We raise these points to observe that even if we can identify effective and appropriate hypothesis tests for typical offline evaluation metrics, that does not fully address the goal of inferring whether or not a proposed system is actually more effective; additional sources of bias need to be accounted for, and it is not clear that NHST is the best framework for evaluating results.

5. Challenges of Statistical Inference and Next Steps

It is important that statistical inference results are reported with all necessary details in order to make research papers as informative as possible, and to estimate and give the reader confidence in understanding the credibility and impact of a reported improvement. However, this is not the prevalent current practice in the RecSys community, as demonstrated by the results in Section 2. While there has been significant attention paid to other aspects of the evaluation

process [2, 3, 4, 5], and the IR community has studied inference for certain experimental settings (see Section 3), this aspect has not yet been a noticeable part of the scholarly discourse on evaluation practices. We argue that this gap needs to be filled.

As a first step, we propose that researchers should report clearly how they performed inference on their results, with multiple results as appropriate. For example, studies using frequentist significance testing should report the test used, the p-value threshold, any corrections for multiple comparisons, and also the effect size and confidence interval, in order to help readers fully understand and better apply the findings. Reporting effect size and sample size help to make papers as informative as possible. While further research is needed to identify best practices for selecting and applying techniques, research using current practices should clearly document them.

We believe further research is needed to identify best practices for applying and reporting on classical tests and techniques, and to study how more advanced inference techniques may be able to mitigate some of their limitations. One such advanced technique that could be studied in the recommender system context is the mixed effect model for testing significance of effects, or Bayesian inference techniques for computing and summarizing posterior distributions of effect sizes. We also believe that, as the community continues work towards documented best practices, and has discussed in the past the need to lay out recommended methods for the benefit of authors, reviewers, and editors [1], such practices need to include recommendations for statistical techniques. The community may be ready to make some such recommendations now, but we call for further research to provide empirical evidence for the appropriateness of recommended techniques, and for such guidelines to leave the door open for innovation in statistical analysis of recommender system evaluations, at least so long as the direction of this innovation is towards greater understanding and rigor.

Acknowledgments

References

[1] J. A. Konstan, G. Adomavicius, Toward identification and adoption of best practices in algorithmic recommender systems research, in: Proceedings of the international workshop on Reproducibility and replication in recommender systems evaluation, 2013, pp. 23–28.

[2] M. F. Dacrema, P. Cremonesi, D. Jannach, Are we really making much progress? a worrying analysis of recent neural recommendation approaches, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 101–109.

[3] S. Rendle, L. Zhang, Y. Koren, On the difficulty of evaluating baselines: A study on recommender systems, arXiv preprint arXiv:1905.01395 (2019).

[4] R. Cañamares, P. Castells, On target item sampling in offline recommender system evaluation, in: Fourteenth ACM Conference on Recommender Systems, 2020, pp. 259–268.

[5] Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, C. Geng, Are we evaluating rigorously?

benchmarking recommendation for reproducible evaluation and fair comparison, in: Fourteenth ACM Conference on Recommender Systems, 2020, pp. 23–32.

[6] G. Shani, A. Gunawardana, Evaluating recommendation systems, in: Recommender systems handbook, Springer, 2011, pp. 257–297.

[7] T. Sakai, 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, pp. 5–14.

[8] M. D. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007, pp. 623–632.

[9] J. Urbano, H. Lima, A. Hanjalic, 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, pp. 505–514.

[10] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, A. Barreiro, Using score distributions to compare statistical significance tests for information retrieval evaluation, Journal of the Association for Information Science and Technology 71 (2020) 98–113.

[11] J. Urbano, M. Corsi, A. Hanjalic, How do metric score distributions affect the type i error rate of statistical significance tests in information retrieval?, in: Conference on the Theory of Information Retrieval (ICTIR'21), 2021.

[12] J. Parapar, D. E. Losada, Á. Barreiro, 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, 2021, pp. 655–664.

[13] M. D. Ekstrand, V. Mahant, Sturgeon and the cool kids: Problems with random decoys for top-n recommender evaluation, in: The Thirtieth International Flairs Conference, 2017.

[14] M. Tian, M. D. Ekstrand, Estimating error and bias in offline evaluation results, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 2020, pp. 392–396.

[15] R. Cañamares, P. Castells, A probabilistic reformulation of memory-based collaborative filtering: Implications on popularity biases, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 215–224.

[16] R. Cañamares, P. Castells, Should i follow the crowd? a probabilistic analysis of the effectiveness of popularity in recommender systems, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 415–424.

[17] A. Bellogín, P. Castells, I. Cantador, Statistical biases in information retrieval metrics for recommender systems, Information Retrieval Journal 20 (2017) 606–634.

[18] B. B. McShane, D. Gal, A. Gelman, C. Robert, J. L. Tackett, Abandon statistical significance, The American Statistician 73 (2019) 235–245.

[19] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al., Redefine statistical significance, Nature human behaviour 2 (2018) 6–10.

[20] T. Sakai, Statistical reform in information retrieval?, in: ACM SIGIR Forum, volume 48, ACM New York, NY, USA, 2014, pp. 3–12.

[21] B. A. Carterette, Multiple testing in statistical analysis of systems-based information retrieval experiments, ACM Transactions on Information Systems (TOIS) 30 (2012) 1–34.