Multiple Testing for IR and Recommendation System Experiments***

Ngozi Ihemelandu[1][0000−0002−8468−1581] and Michael D. Ekstrand[2][0000−0003−2467−0108]

[1] Boise State University, Boise ID 83725, USA ngoziihemelandu@u.boisestate.edu
[2] Dept. of Information Science, Drexel University, Philadelphia PA 19104, USA ekstrand@acm.org

Abstract. While there has been significant research on statistical techniques for comparing two information retrieval (IR) systems, many IR experiments test more than two systems. This can lead to inflated false discoveries due to the multiple-comparison problem (MCP). A few IR studies have investigated multiple comparison procedures; these studies mostly use TREC data and control the familywise error rate. In this study, we extend their investigation to include recommendation system evaluation data as well as multiple comparison procedures that controls for False Discovery Rate (FDR).

Keywords: statistical Inference · family-wise error rate · false discovery rate · multiple testing · recommender system · information retrieval

1 Introduction

Effective evaluation of information retrieval and recommender systems requires an assessment of whether the difference in performance metrics observed in an experiment likely represent a real improvement. Statistical tests are designed to fill this gap, assessing the likelihood that an observed improvement could be seen with random chance. Several studies have investigated which significance tests are most appropriate for analyzing evaluation results [15, 16, 13, 10, 19, 21, 14, 18, 12] mostly examining comparisons between two systems; a few studies [6, 20, 5] consider comparing more than two systems.

Comparing all pairs of k systems requires $m = k(k-1)/2$ tests, while the significance level α controls the probability of falsely finding significance only for a single test; with m tests, the probability of incorrectly finding a significant

* Partly supported by the National Science Foundation on Grant 17-51278.

difference increases to $1 - (1 - \alpha)^m$. This is known as the multiple comparison problem (MCP).

One common approach to addressing the MCP is to adjust the p-value to control the traditional family-wise error rate (FWER). FWER is the probability of making at least one false positive in m experiments. Another approach to adjusting p-values is to control the false discovery rate (FDR) [2]. FDR is the expected proportion of false positive results out of all positive test results.

When we fail to apply appropriate statistical analysis that addresses the MCP, we run the risk of either failing to advance algorithmic methods that should be advanced or falsely identifying interesting results where none exist. Both of these problems can hinder progress in information retrieval research and application both by holding back improvements and by spending time on methods whose observed improvement was a fluke.

Previous IR studies that have investigated the MCP focused on multiple comparison methods that controlled the FWER and used TREC data for their analysis. In this study, we extend our analysis to include procedures that control the FDR in p-value adjustment and to use recommendation system evaluation data. Recommendation system data have a few distinguishing features such as large sample size and high sparsity which induces bias in effectiveness metrics [1]. Our goal is to understand how controlling for each of these different error rate impacts IR and recommendation system evaluation result analysis and enable researchers choose the appropriate test for their analysis.

To that end, we address the following research questions:

- **RQ1** - When systems have equivalent performance (the null hypothesis is true) do procedures that correct for MCP control the family-wise error rate or false discovery rate at a specified α level?
- **RQ2** - How many missed findings does controlling for FWER and FDR lead to respectively?
- **RQ3** - Which method does best in identifying as many actual differences between systems as possible while still maintaining a low false positive rate?

We find that multiple correction tests meets its objective of controlling the FWER and FDR at or below the given α level when the sample size is small (≤ 1000). However, when both the sample size and the number of hypothesis tests are large, they control the error rate at a much higher level than the target α level. We also observe that correction tests that control the false discovery rate instead of the family-wise error rate are more powerful.

2 Related Works and Background

A number of studies [12, 6, 21, 14] have investigated appropriate statistical methods for analyzing IR evaluation results for experiment comparing two systems. A few studies have focused on multiple testing adjustments when comparing more than two systems. Tague-Sutcliffe and Blustein [20] adjusted the p-value using the Scheffe's method [17] and found that only large effect sizes could be detected.

Boytsov et al. [5] focused on adjusting p-values for non-parametric statistical procedures and found that the correction procedures found fewer true differences compared to unadjusted tests. Carterette [6] used a single-step method that relied on multivariate Student distribution to adjust the p-value for MCP; like [20] found that small pairwise differences were not detected.

There are two main approaches used in multiple hypothesis testing to correct for multiple comparisons or alpha inflation: controlling for family-wise error rate and controlling for false discovery rate.

2.1 Controlling the family-wise error rate

The family-wise error rate is the probability of having one or more false positives out of all the hypothesis tests conducted. To guarantee that the probability of having one or more false positives in m tests is α or less, Bonferonni adjusts the significance level of each hypothesis test to α/m. We consider two of the popular procedures in this study: Bonferroni [4] and Holms [9].

2.2 Controlling the false discovery rate

The false discovery rate (FDR) is the expected proportion of false positives (or discoveries) among all positives/discoveries. Controlling the FDR instead of the FWER is a more recent approach to addressing the multiple comparison problem. When all hypothesis are true, this error rate is equivalent to the FWER [2] but may not be controlled at the same level otherwise. We investigated the Benjamini & Hochberg (BH) [2] and Benjamini-Yekuteili (BY) [3] multiple correction procedures in this study. BY makes fewer assumptions than BH.

3 Methodology and Data

We study the behavior of multiple-comparison corrections using simulated replications of search experiments on TREC 2013 Web [7] and a recommendation experiment with MovieLens 100K [8]. We adopt the methodology developed by Urbano and Nagler [22], which is based on marginal distributions and inter-system dependencies. This methodology employs evaluation scores from each algorithm run to construct parametric and semi-parametric models. These models represent the marginal distribution of per-topic effectiveness scores for a system, as well as vine copulas that models inter-system dependencies.

The TREC data comes with runs, and for MovieLens, we generated runs using four collaborative filtering recommender algorithms (ALS BiasedMF, ALS ImplicitMF, Item k-NN, and User-based k-NN) in various configurations from the LensKit toolkit. All runs are evaluated using normalized Discounted Cumulative Gain (nDCG), with a cutoff threshold set at 20 for TREC data and 100 for recommender system data.

This combination generates data that is realistic both in terms of score distributions and inter-system correlations. To simulate an experiment with k systems, the process begins by fitting a model to generate replicates:

1. Randomly select k systems and their runs from the original data set.
2. Fit a marginal distribution F_B to each run B. Using code from Urbano and Nagler, candidate distributions take parametric (Truncated Normal, Beta, and Beta-Binomial) and non-parametric (discrete kernel smoothing) forms, and the best-fitting distribution is used.
3. Fit a Vine copula to model the joint score distributions between runs.

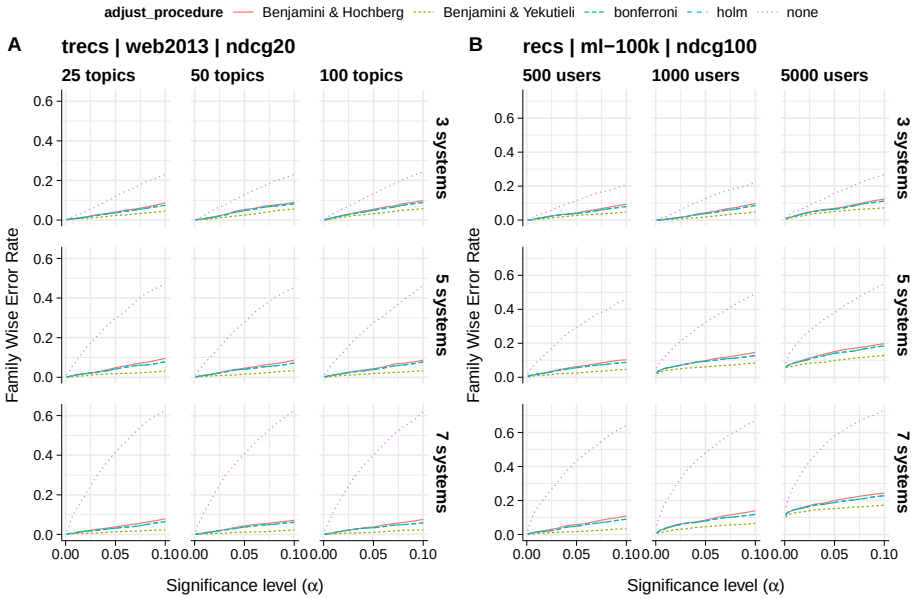

Fig. 1. When systems are equivalent (null hypothesis is true), the error rate is controlled at a significance level (α) by each adjustment test. The control of the error rate is determined by the sample size and number of hypothesis tests. We use all pairwise hypothesis test which determines the number of hypothesis tests. (For example, 5 systems would give 10 hypothesis tests).

We assess significance with pairwise t-tests between systems [11]. To address **RQ1**, which considers scenarios where systems have equivalent performance (thus making the null hypothesis true), we simulate true null hypotheses with all compared systems having the same mean effectiveness. The process involves:

1. randomly selecting a run, denoted as k_0, from the k runs.
2. calculating its mean effectiveness μ_0.
3. transforming the marginal distributions of the remaining $k-1$ runs to have the mean μ_0.

To address **RQ2**, which examines the ability of procedures controlling for FWER and FDR to detect effects when they exist (i.e., when the null hypothesis is false),

we simulate scenarios where system E outperforms system B by a specific effect size of δ (effect size δ quantifies the magnitude of the difference in effectiveness between two recommender algorithms, providing insight into the practical or real-world significance of the findings). The simulation steps are as follows:

1. randomly select a run, denoted as k_0 from the k runs.
2. calculate its mean effectiveness, μ_0.
3. transform the marginal distributions of the other runs to have a mean of $\mu_0 + \delta$.

To address **RQ3**, which aims to identify the correction method most effective at detecting actual differences between systems while maintaining a low false positive rate (in scenarios with a mix of true and false null hypotheses), we simulate a mixture of equal and non-equal systems. The simulation process involves:

1. selecting a pair of runs and transforming their marginal distributions to have the same mean, μ_0.
2. choosing a different pair of runs and adjusting their marginal distributions so that the difference in their means is δ.

We then simulate new experimental data with a sample size n:

1. Adjust marginal distributions for selected runs to match experimental condition (all systems equal, one system better and $k-1$ equal, or a mix of equal and non-equal systems).
2. Draw n topics each with k pseudo-observations from the fitted copula.
3. Apply the adjusted inverse Cumulative Distribution Function (CDF) F_B^{-1} of each system to the corresponding pseudo-observations to get final scores.
4. Use paired t-test to compare all pairs of systems.
5. Correct the t-test's p-values using the selected multiple comparison correction procedure to yield corrected p-values.
6. Calculate the false positives and/or power (as applicable) of comparing the adjusted p-value to the significance threshold α.

We ran 10,000 simulations for each configuration, using $k \in \{3, 5, 7\}$ and n ($n \in \{25, 50, 100\}$ for TREC, $n \in \{500, 1000, 5000\}$ for MovieLens). For the case of one system outperforming several baselines, we used effect sizes of 0.01, 0.05, and 0.1; for a mix, we had pairs of equivalent systems and individual systems with effect size δ over the pairs.

4 Results

Figure 1 shows the case corresponding to **RQ1**, wherein the null hypothesis is true, indicating that all systems exhibit equivalent performance. We observe that when the number of systems and sample size are small, or the number of systems is large and the sample size is small, the correction procedures control the error rate at (or below) the specified α significance level. However, when

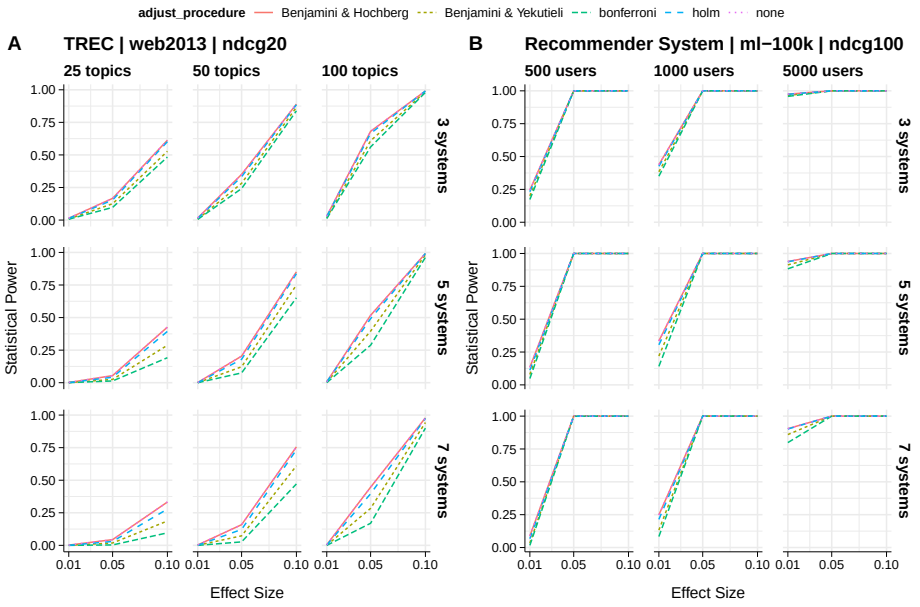

Fig. 2. The ability of adjustment procedures to find 'real' effects (statistical power) while controlling for FWER and FDR at $\alpha = 0.05$.

the sample size is large and the number of systems is large (specific with the recommendation dataset), the error rate is inflated, but not nearly as inflated as the uncorrected test.

Figure 2 presents the results for **RQ2**, showing that with the TREC style experiment, when the sample size is small and when many tests are being conducted, the correction procedures impose a fairly severe penalty which reduces the statistical power. This is more pronounced with the Bonferroni test. Holms and Benjamini & Hochberg tests have the most power while the Bonferroni test has the least power. However, with the recommendation experiment, we observe differences in statistical power between the adjustment tests only when both the sample size and effect size are small. For the large sample sizes, there are no differences in the statistical power between the correction procedures.

Figure 3 displays the results for **RQ3**, highlighting the influence of sample size and the number of systems on the balance between error rates and the power of correction procedures. We observe that as the sample size and number of systems increase, the error rate and power also increase, while power decreases and the error rate is maintained at or below the specified threshold when the number of systems increases but the sample size is small.

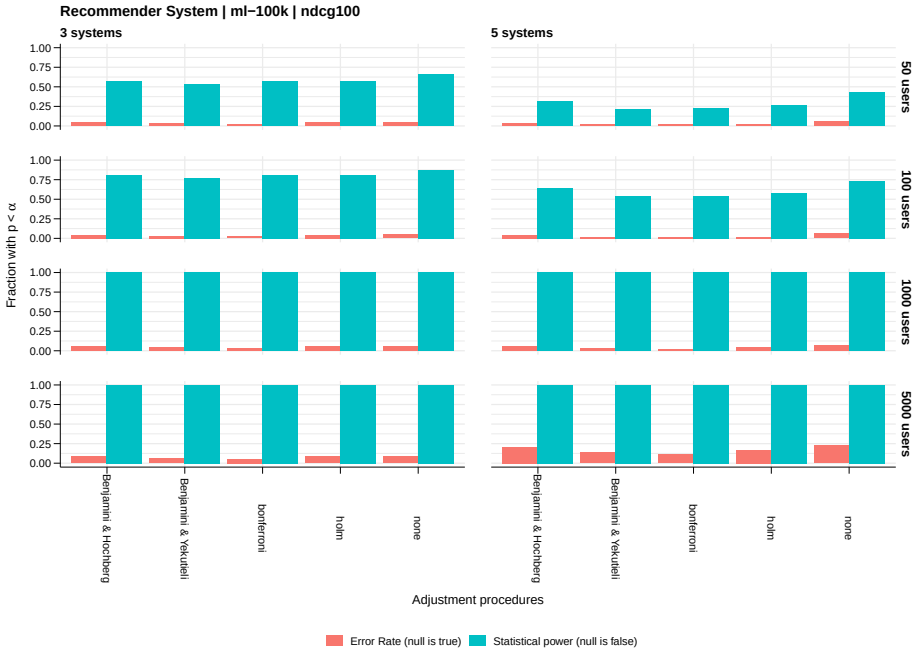

Fig. 3. A mixture of null hypothesis and non-null hypothesis (with effect size = 0.05). The proportion of null hypothesis that are $\leq \alpha$ (error rate) and the proportion of non-null hypothesis that are $< \alpha$ (statistical power).

5 Discussion and Conclusion

We find that corrections for multiple comparisons generally behave as expected in the small sample sizes of TREC-style IR experiments, controlling the error rate below the target significance threshold. However, as the sample size and number of systems increases, as in a large-scale search or recommendation experiment, the corrections no longer keep the overall error rate under the target level, although they still result in far fewer false discoveries than uncorrected pairwise t-tests. Our results also show that while the corrections have differing power and fail to find small effects in small experiments, they recover their power in a larger-scale experiment, and small effects are easily found even with conservative corrections like Bonferonni.

The Benjamini-Yekuteili test showed the lowest error rate in experiments with all systems equivalent, had greater power than the Bonferonni test at small to medium sample sizes, and strikes a balance between power and error in mixed-effect-size experiments. We therefore recommend it as the default correction for comparing multiple systems in an IR experiment. In large-scale experiments, if the computational and conceptual simplicity of the Bonferonni test is preferred, it can be used without meaningful loss in power.

References

1. Bellogín, A., Castells, P., Cantador, I.: Statistical biases in information retrieval metrics for recommender systems. Information retrieval journal **20**, 606–634 (2017)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
3. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of statistics pp. 1165–1188 (2001)
4. Bland, J.M., Altman, D.G.: Multiple significance tests: the bonferroni method. Bmj **310**(6973), 170 (1995)
5. Boytsov, L., Belova, A., Westfall, P.: Deciding on an adjustment for multiplicity in ir experiments. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. pp. 403–412 (2013)
6. Carterette, B.A.: Multiple testing in statistical analysis of systems-based information retrieval experiments. ACM Transactions on Information Systems (TOIS) **30**(1), 1–34 (2012)
7. Hagen, M., Völske, M., Gomoll, J., Bornemann, M., Ganschow, L., Kneist, F., Sabri, A.H., Stein, B.: Webis at trec 2013-session and web track. In: TREC (2013)
8. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis) **5**(4), 1–19 (2015)
9. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics pp. 65–70 (1979)
10. Hull, D.: 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. pp. 329–338 (1993)
11. Ihemelandu, N., Ekstrand, M.D.: Statistical inference: The missing piece of recsys experiment reliability discourse. arXiv preprint arXiv:2109.06424 (2021)
12. Ihemelandu, N., Ekstrand, M.D.: 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) (2023)
13. Jones, K.S., Willett, P.: Readings in information retrieval. Morgan Kaufmann (1997)
14. Parapar, J., Losada, D.E., Presedo-Quindimil, M.A., Barreiro, A.: Using score distributions to compare statistical significance tests for information retrieval evaluation. Journal of the Association for Information Science and Technology **71**(1), 98–113 (2020)
15. Rijsbergen, C.v.: Van. Information Retrieval, Butterworths **2** (1979)
16. Savoy, J.: Statistical inference in retrieval effectiveness evaluation. Information Processing & Management **33**(4), 495–512 (1997)
17. Scheffé, H.: A method for judging all contrasts in the analysis of variance. Biometrika **40**(1-2), 87–110 (1953)
18. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on conference on information and knowledge management. pp. 623–632 (2007)
19. Tague-Sutcliffe, J.: The pragmatics of information retrieval experimentation, revisited. Information processing & management **28**(4), 467–490 (1992)

20. Tague-Sutcliffe, J., Blustein, J.: A statistical analysis of the trec-3 data. NIST Special Publication SP pp. 385–385 (1995)
21. Urbano, J., Lima, H., Hanjalic, A.: 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. pp. 505–514 (2019)
22. Urbano, J., Nagler, T.: Stochastic simulation of test collections: Evaluation scores. In: The 41st international ACM SIGIR conference on research & development in information retrieval. pp. 695–704 (2018)