Towards Reliable Testing for Multiple Information Retrieval System Comparisons

David Otero Javier Parapar Álvaro Barreiro

IRlab, CITIC, Universidade da Coruña, Spain

UNIVERSIDADE DA CORUÑA

Motivation

- IR evaluation relies on test collections with limited number of topics (e.g., TREC collections with 50 topics).

Motivation

- IR evaluation relies on test collections with limited number of topics (e.g., TREC collections with 50 topics).
- Null Hypothesis Significance Testing (NHST) helps to improve the certainty we can have in results with these collections.

Motivation

- IR evaluation relies on test collections with limited number of topics (e.g., TREC collections with 50 topics).
- Null Hypothesis Significance Testing (NHST) helps to improve the certainty we can have in results with these collections.
- Commonly, IR experiments involve more than two systems. Thus, we test multiple hypotheses at the same time.

Motivation

- IR evaluation relies on test collections with limited number of topics (e.g., TREC collections with 50 topics).
- Null Hypothesis Significance Testing (NHST) helps to improve the certainty we can have in results with these collections.
- Commonly, IR experiments involve more than two systems. Thus, we test multiple hypotheses at the same time.
- **Multiple comparisons increase the chance of false positives.**

The Multiple Comparison Problem

- Comparing m systems requires $k = \frac{m(m-1)}{2}$ pairwise tests.

The Multiple Comparison Problem

- Comparing m systems requires $k = \frac{m(m-1)}{2}$ pairwise tests.
- If the probability of a Type I error is α,

The Multiple Comparison Problem

- Comparing m systems requires $k = \frac{m(m-1)}{2}$ pairwise tests.
- If the probability of a Type I error is α, the probability of making **at least** one Type I error is $1 - (1 - \alpha)^k$ (**this is the FWER!**).

The Multiple Comparison Problem

- Comparing m systems requires $k = \frac{m(m-1)}{2}$ pairwise tests.
- If the probability of a Type I error is α, the probability of making **at least** one Type I error is $1 - (1 - \alpha)^k$ (**this is the FWER!**).
- Unadjusted significance tests lead to high Type I error rates.
  - Studies show 40-50% of uncorrected tests yield false positives. [Ferro and Sanderson. SIGIR '24]

Controlling the FWER

- Several strategies exist to control the FWER.
- The main idea is to make the test more conservative.
- Examples: Bonferroni, Holm's method, ANOVA+TukeyHSD, randomised TukeyHSD, etc.

Controlling the FWER

- Several strategies exist to control the FWER.
- The main idea is to make the test more conservative.
- Examples: Bonferroni, Holm's method, ANOVA+TukeyHSD, randomised TukeyHSD, etc.
- **Too conservative? $\rightarrow$ Less power**

False Discovery Rate (FDR)

- **Alternative to FWER**: control the FDR.
- The FDR is the expected proportion of false discoveries **among the rejected null hypotheses**.
- The rationale is to make the tests less strict but still be able to control the Type I errors.
- Examples: Benjamini-Hochberg (BH), Benjamini-Yekutieli (BY).

This work

- We evaluate the behaviour of several statistical methods for multiple comparisons in IR.

- We use two approaches:
  - **Simulated Data Experiment**: create variations of an IR system with known differences.
  - **Real Data Validation**: use Million Query Track dataset to compute long-term performance over hundreds of topics.

Our simulation (adapted from [Parapar et al., ACM SAC '21])

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

- Use Logistic Regression for modelling the appearance of relevant documents in a ranking.
- We fit a regressor that has the form of (1), where p is the position in the ranking and θ_0 and θ_1 are the fitted parameters.
- We fit a regressor for each topic-system pair. For each system, we can simulate m variations.

Our simulation (adapted from [Parapar et al., ACM SAC '21])

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

- Sampling from this regressor gives a probability h_θ of relevance which we use to sample from a Bernoulli distribution.
- We can manipulate the paramters θ_0 and θ_1 to create situations where H_0 is false.

We simulate two different scenarios

Scenario 1: we sample m times from each regressor without modifying θ_0 and θ_1. Every system is statistically equivalent, and thus **every null hypothesis is true**.

Scenario 2: we sample m times from each regressor while improving θ_0 and θ_1 by a given proportion. **Every null hypothesis is false**.

Experimental Setup

- Statistical Tests:
  - Tests: t-test, Wilcoxon Signed-Rank test.
  - Adjustments: Bonferroni, Holm, Benjamini-Hochberg (BH), Benjamini-Yekutieli (BY).
  - Alternatives: ANOVA with TukeyHSD, Randomised TukeyHSD.

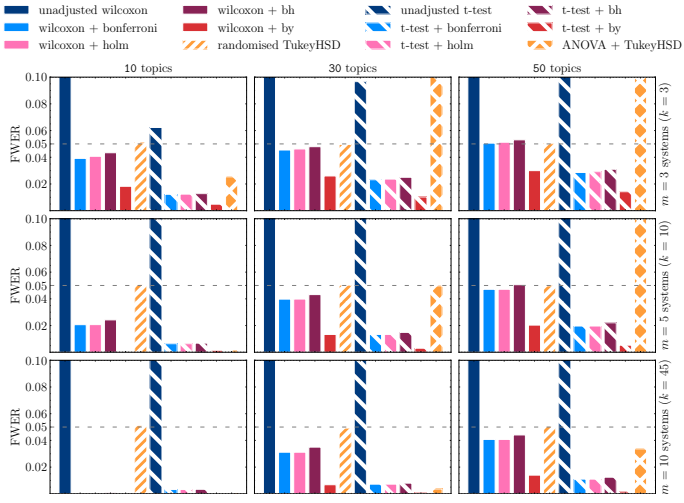
- TREC-8 collection, sample size $t \in \{10, 30, 50\}$, number of systems $m \in \{3, 5, 10\}$

Scenario 1: Type I Error Rate Evaluation

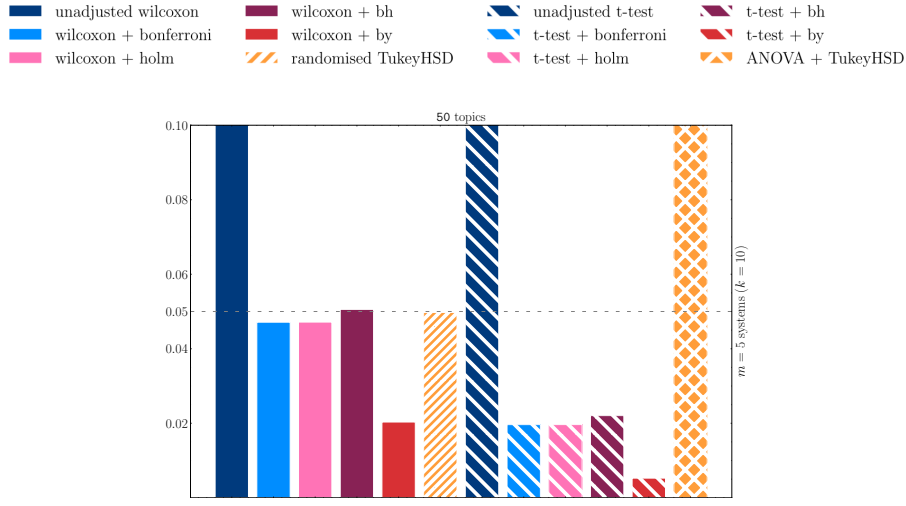
- For each simulation step, we have a family of k null hypotheses, where every hypothesis should not be rejected.

- We compute the FWER as the number of times that a test marked *at least* one hypothesis as significant.

Results: Type I Error Rates



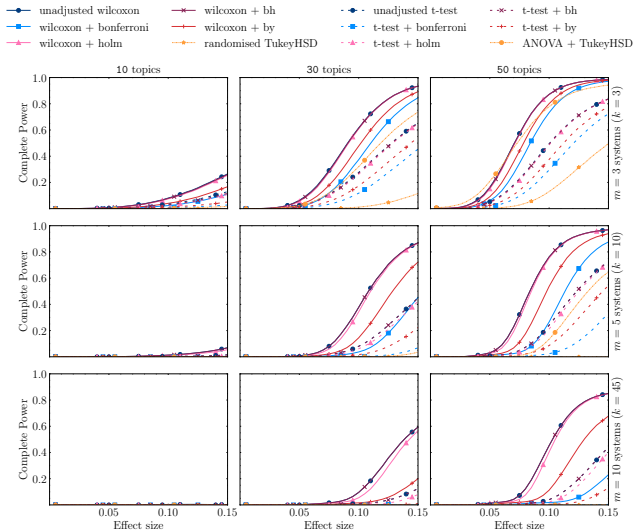
Type I Error Rates for 50 topics and 5 systems



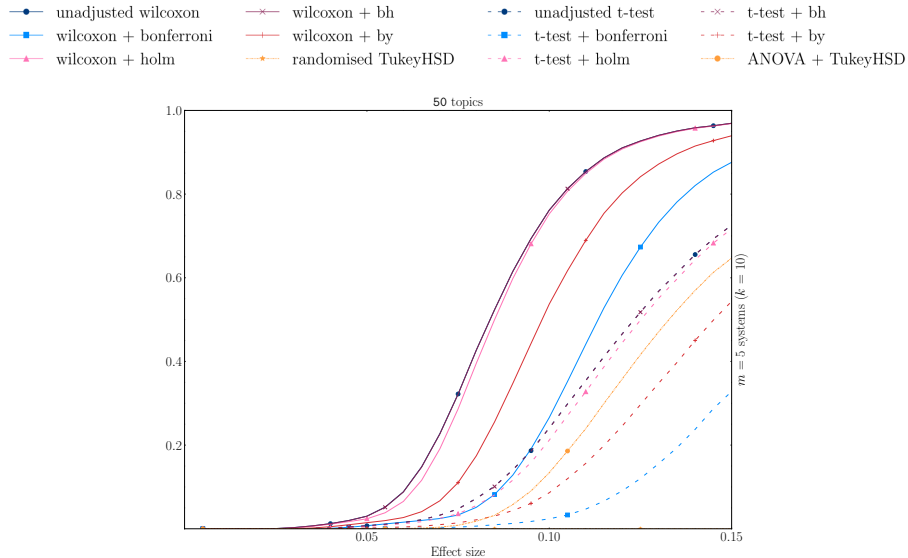
Scenario 2: Power Evaluation

- For each simulation step, we have a family of k null hypotheses, where every hypothesis should be rejected.
- We compute the Complete Power as the number of times that a test marked *every* hypothesis as significant.

Results: Power



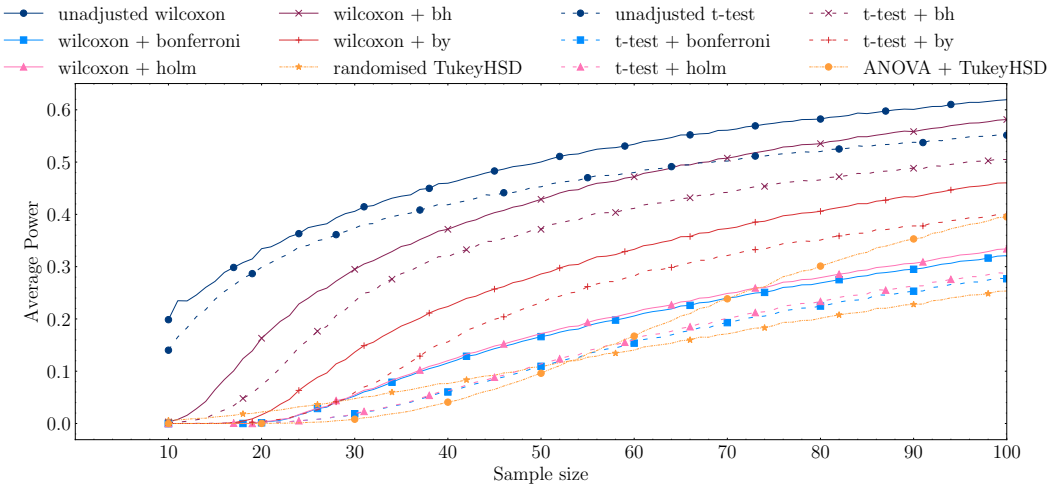
Power for 50 topics and 5 systems



Experiment with Million Query Track Data
adapted from [Boytsov et al., SIGIR '13]

1. Compute average scores over 687 topics, We assume that averages over hundreds of topics are representative of long-term performance, and mark as different those pairs that have a MAP difference higher than 0.05%.
2. Sample random subsets of these topics, and **compare if the outputs of the tests on these subsets agree with the long-term performance**.

Million Query Results



Conclusions

- Proper statistical adjustments are crucial in IR evaluation.
- Tests that do not match the expected error rate perform poorly in terms of power.
- Wilcoxon + BH provides the best trade-off between Type I error rate and power.

Towards Reliable Testing for Multiple Information Retrieval System Comparisons

David Otero Javier Parapar Álvaro Barreiro

IRlab, CITIC, Universidade da Coruña, Spain

UNIVERSIDADE DA CORUÑA