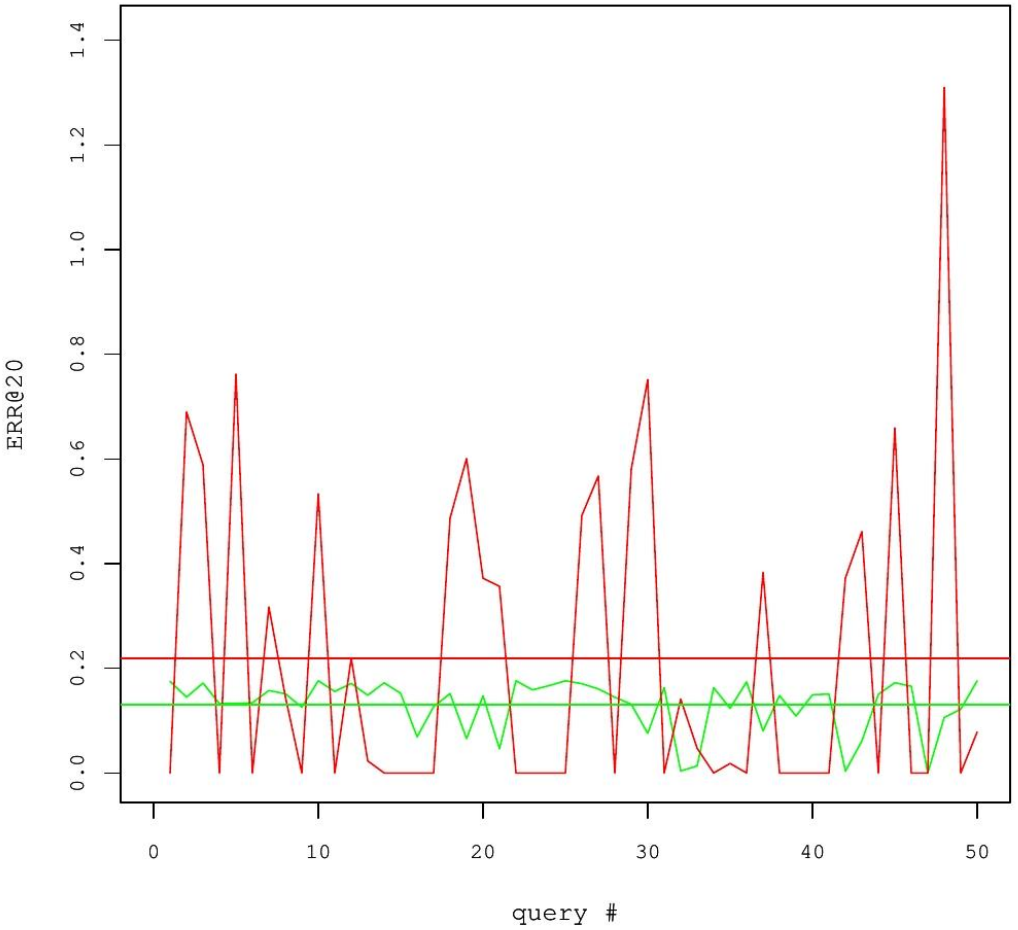
Deciding on Multiplicity Adjustments in IR Experiments

Leonid Boytsov (Carnegie Mellon University)
Anna Belova (Abt Associates Inc)
Peter Westfall (Texas Tech University)

Possible comparison scenario

- Evaluate performance scores for each query and a system
- Compute averages for each system
- Compare average performance scores of systems

Why Statistical Testing?



Paired Difference Test

$$X = \left(X_1, X_2, \ldots, X_q\right)$$

$$Y = \left(Y_1, Y_2, \ldots, Y_q\right)$$

H_0: $$E\left[\frac{1}{q}\sum X_i - Y_i\right] = 0$$

Which Statistical Test?

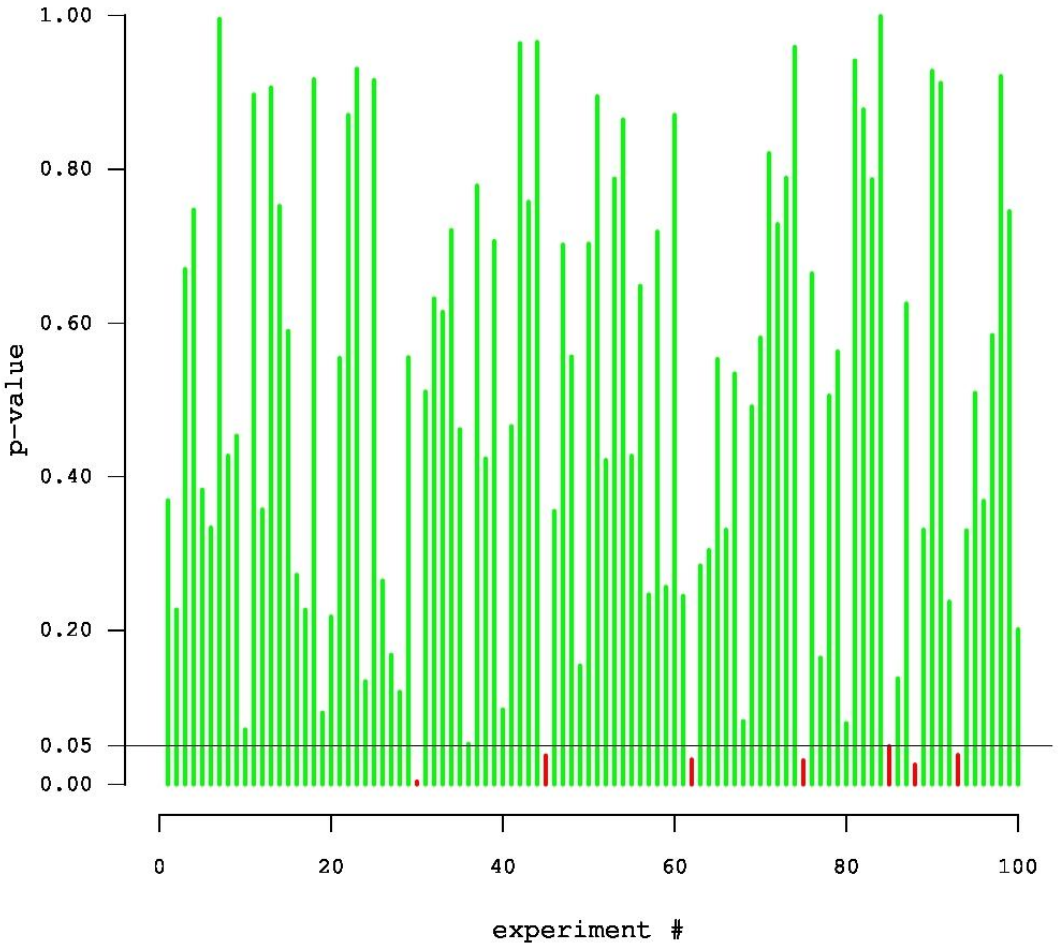
Test	Assumptions
Student's paired *t*-test	$X_i - Y_i$ are i.i.d. and normal
Wilcoxon signed-rank	$X_i - Y_i$ are i.i.d. and symmetric around zero
Sign test	$X_i - Y_i$ are i.i.d.
Bootstrapping	$X_i - Y_i$ are i.i.d.
Permutation	Query-level exchangeability of X and Y. $X_i - Y_i$ are not necessarily i.i.d.

Which Statistical Test?

- Bootstrap, permutation, and t-test are equally good[1]
- Wilcoxon and Sign test can be inferior[2]

[1] Wilbur 1994, Smucker et al 2007, our experiments
[2] Wilbur 1994, Smucker et al 2007

What about Multiple Tests?



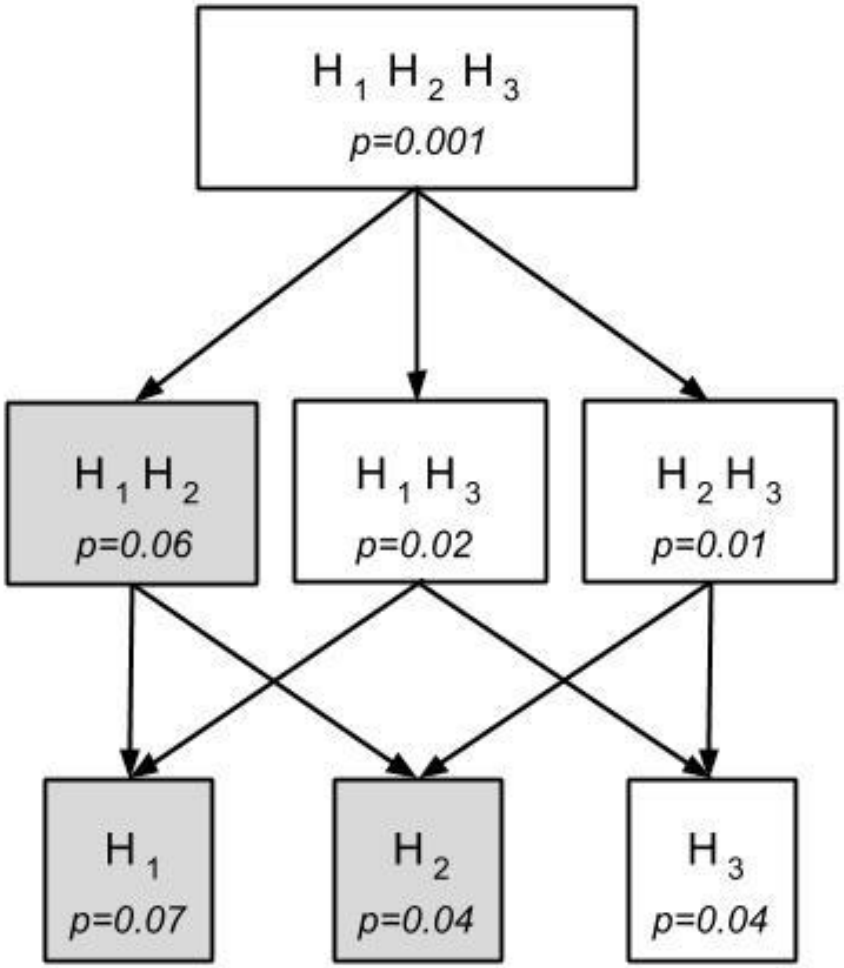
How To Fix: Adjust p-Values

- Classic Bonferroni adjustment (multiply p-values by the # of experiments)
- Holm-Bonferroni adjustment (marginally improves on Bonferroni)

Problems?

- Considered to be conservative
- Correlations are not taken into account
- Example: the same system compared against the same baseline 100 times (using the same test collection)

Closed Testing is Better!



Closed Testing is Better!

- Recall the example where we repeat the same test 100 times (perfect correlation)
- All joint null hypotheses are same:

$$H_0 = H_0 \bigcap H_0 \bigcap \ldots \bigcap H_0$$

- The p-value obtained for each joint hypothesis can be (roughly) the same as in a single test

But Much more Expensive…

$2^{\#\text{ of hypotheses}}$

But Much more Expensive…

$2^{\#}$ of hypotheses

Super-slow if joint hypotheses are tested using permutation

But Much more Expensive…

$2^{\#}$ of hypotheses

Super-slow if joint hypotheses are tested using permutation

There is an efficient **approximation**:

MaxT permutation test

Need to Test in IR Setting

- Unadjusted p-values (permutation)
- Holm-Bonferroni correction
- Closed Testing (permutation)
- MaxT permutation test

TREC data

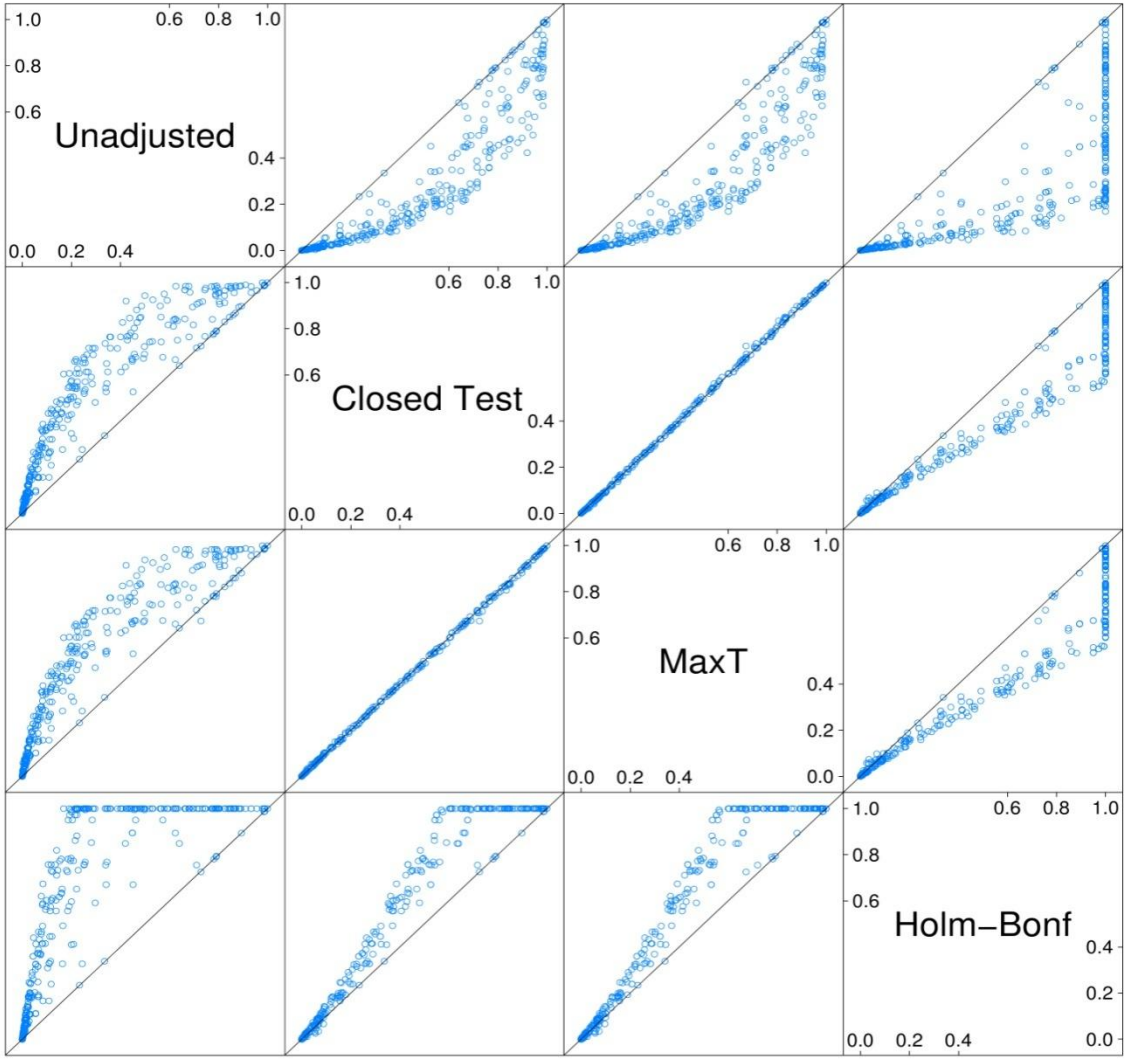
- Real runs (ad hoc, 50 queries)
- No ground truth performance information

Losses Due to Adjustment[1]

TREC	Closed Test	MaxT	Holm Bonf
3	16.1%	16.4%	19.1%
4	12.7%	12.7%	15.4%
5	7.5%	8.7%	10%
7	15%	15.4%	17.3%
8	8.2%	8.2%	9.5%
19	31.1%	32.1%	32.1%
20	33.5%	33.5%	38.1%
All	**16.4%**	**16.8%**	**18.8%**

[1] As a fraction of the number of significant results (without adjustment for multiplicity)

Agreement in p-values (TREC)



Microsoft Learning To Rank (MSLR)

- Synthesize realistic 30K-query runs
- Some of them had same performance
- ERR@10 in the 30K-query population represent the ground truth performance

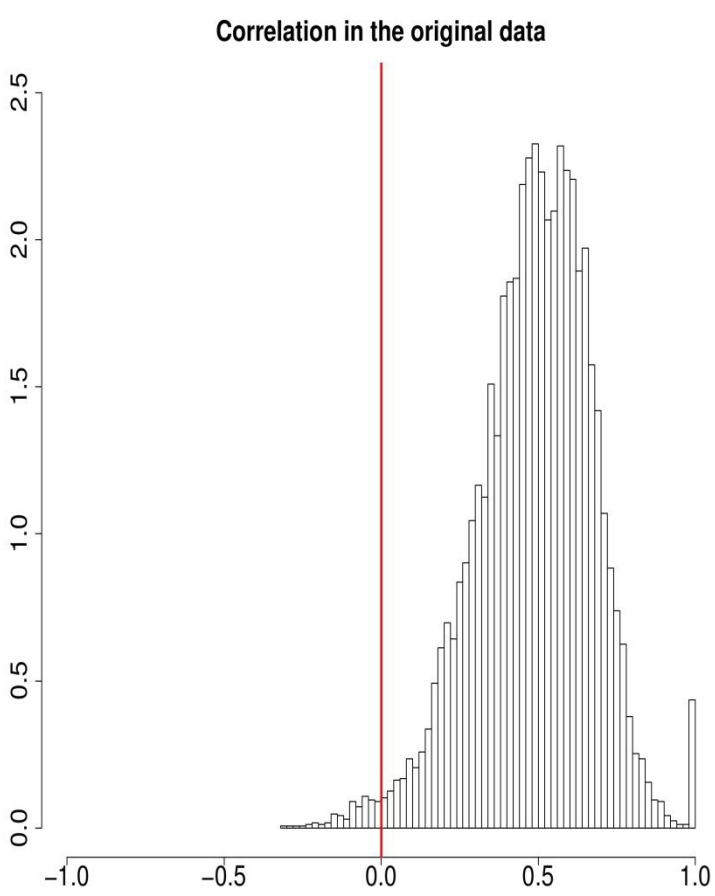
Microsoft Learning To Rank (MSLR)

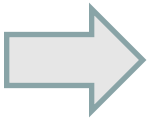
- Create shorter runs by sampling from 30K queries
- Compare runs using our statistical tests
- Compute the number of false positives/negatives

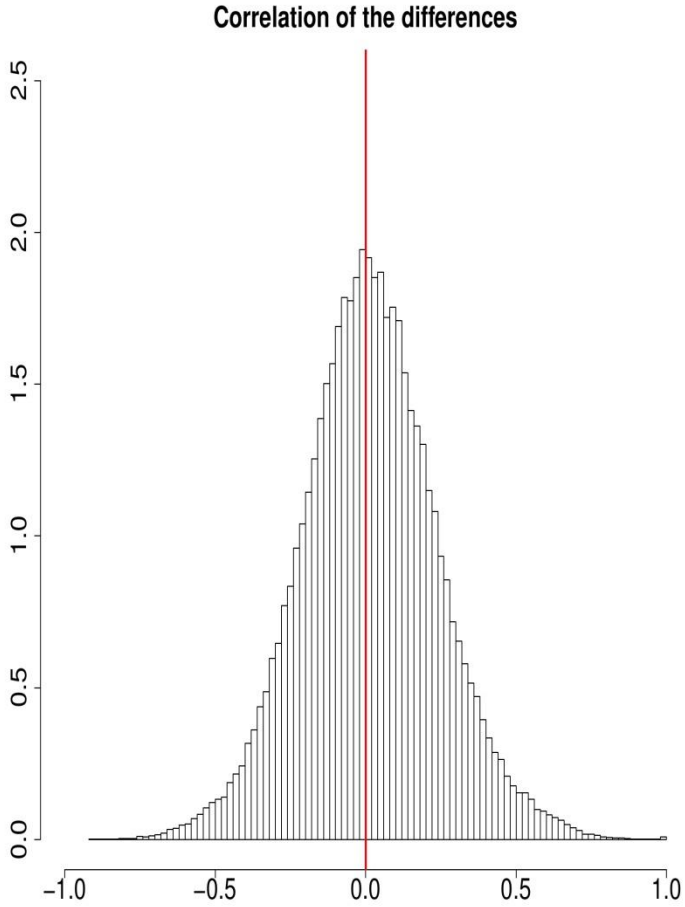
Results (MSLR)

Query set size					
	50	100	400	1600	6400
Language Models Runs (similar to real TREC systems)					
Unadjusted	**85.7/14.4**	**80.8/11.6**	**53.9/10.0**	**25.9/15.4**	**2.5/17.0**
Closed test	**92.9/0.0**	**88.8/0.2**	**69.5/1.7**	**36.6/3.1**	**5.2/6.8**
MaxT	**93.9/0.0**	**91.8/0.2**	**68.0/1.2**	**35.7/3.0**	**6.3/6.6**
Holm-Bonf.	**94.9/2.0**	**92.5/1.8**	**69.6/2.6**	**37.0/3.2**	**6.5/6.2**
Format: false negatives (**dark blue**)/false positives (**red**)					

Correlations?







Summary

We recommend MaxT, because:

1. Detection rates can be good (especially with large query samples)
2. MaxT is efficient and the p-values are nearly identical to those of inefficient Closed Testing
3. Unlike Holm-Bonferroni, MaxT cannot be “fooled” by perfectly correlated tests

Implementation

https://github.com/searchivarius/PermTest

The $64,000 Question

To improve detection rates, can we still leverage correlations in data when correlations in differences are small?

Simplified Permutation Example[1]

$$\mathrm{SampleMeanDiff}(X, Y) = 0$$

X	**0.13**	**0.14**
Y	**0.15**	**0.12**
Z	**0.16**	**0.11**

[1]In our work t-statistic is used

Simplified Permutation Example[1]

$$\mathrm{SampleMeanDiff}(X, Y) = 0.02$$

X	**0.15**	**0.14**
Y	**0.13**	**0.12**
Z	**0.16**	**0.11**

[1]In our work t-statistic is used

Simplified Permutation Example[1]

$$\mathrm{SampleMeanDiff}\left(X, Y\right) = 0.005$$

X	**0.13**	**0.14**
Y	**0.15**	**0.11**
Z	**0.16**	**0.12**

[1]In our work t-statistic is used