

# Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments

BENJAMIN A. CARTERETTE

University of Delaware

---

High-quality reusable test collections and formal statistical hypothesis testing have together allowed a rigorous experimental environment for information retrieval research. But as Armstrong et al. [2009] recently argued, global analysis of those experiments suggests that there has actually been little real improvement in ad hoc retrieval effectiveness over time. We investigate this phenomenon in the context of simultaneous testing of many hypotheses using a fixed set of data. We argue that the most common approach to significance testing ignores a great deal of information about the world, and taking into account even a fairly small amount of this information can lead to very different conclusions about systems than those that have appear in published literature. This has major consequences on the interpretation of experimental results using reusable test collections: it is very difficult to conclude that *anything* is significant once we have modeled many of the sources of randomness in experimental design and analysis.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software—*Performance Evaluation*

General Terms: Experimentation, Measurement, Theory

Additional Key Words and Phrases: information retrieval, effectiveness evaluation, test collections, experimental design, statistical analysis

---

## 1. INTRODUCTION

The past 20 years have seen a great improvement in the rigor of information retrieval experimentation, due primarily to two factors: high-quality, public, portable test collections such as those produced by TREC (the Text REtrieval Conference [Voorhees and Harman 2005]), and the increased practice of statistical hypothesis testing to determine whether measured improvements can be ascribed to something other than random chance. Together these create a very useful standard for reviewers, program committees, and journal editors; work in information retrieval (IR) increasingly cannot be published unless it has been evaluated using a well-constructed test collection and shown to produce a statistically significant improvement over a good baseline.

But, as the saying goes, any tool sharp enough to be useful is also sharp enough to be dangerous. The outcomes of significance tests are themselves subject to random chance: their  $p$ -values depend partially on random factors in an experiment, such as the topic sample, violations of test assumptions, and the number of times a par-

---

...

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© 20YY ACM 1046-8188/YY/00-0001 \$5.00

ticular hypothesis has been tested before. If publication is partially determined by some selection criteria (say,  $p < 0.05$ ) applied to a statistic with random variation, it is inevitable that ineffective methods will sometimes be published just because of randomness in the experiment. Moreover, repeated tests of the same hypothesis exponentially increase the probability that at least one test will incorrectly show a significant result. When many researchers have the same idea—or one researcher tests the same idea enough times with enough small algorithmic variations—it is practically guaranteed that significance will be found eventually. This is the *Multiple Comparisons Problem*. The problem is well-known, but it has some far-reaching consequences on how we interpret reported significance improvements that are less well-known. The biostatistician John Ioannidis [2005a; 2005b] argues that one major consequence is that most published work in the biomedical sciences should be taken with a large grain of salt, since significance is more likely to be due to randomness arising from multiple tests across labs than to any real effect. Part of our aim with this work is to extend that argument to experimentation in IR.

Multiple comparisons are not the only sharp edge to worry about, though. Reusable test collections allow repeated application of “treatments” (retrieval algorithms) to “subjects” (topics), creating an experimental environment that is rather unique in science: we have a great deal of “global” information about topic effects and other effects separate from the system effects we are most interested in, but we ignore nearly all of it in favor of “local” statistical analysis. Furthermore, as time goes on, we—both as individuals and as a community—learn about what types of methods work and don’t work on a given test collection, but then proceed to ignore those learning effects in analyzing our experiments.

Armstrong et al. [2009b] recently noted that there has been little real improvement in ad hoc retrieval effectiveness over time. That work largely explores the issue from a sociological standpoint, discussing researchers’ tendencies to use below-average baselines and the community’s failure to re-implement and comprehensively test against the best known methods. We explore the same issue from a theoretical statistical standpoint: what would happen if, in our evaluations, we were to model randomness due to multiple comparisons, global information about topic effects, and learning effects over time? As we will show, the result is that we actually *expect* to see that there is rarely enough evidence to conclude significance. In other words, even if researchers were honestly comparing against the strongest baselines and seeing increases in effectiveness, those increases would have to be very large to be found truly significant when all other sources of randomness are taken into consideration.

This work follows previous work on the analysis of significance tests in IR experimentation, including recent work in the TREC setting by Smucker [2009; 2007], Cormack and Lynam [2007; 2006], Sanderson and Zobel [2005], and Zobel [1998], and older work by Hull [1993], Wilbur [1994], and Savoy [1997]. Our work approaches the problem from first principles, specifically treating hypothesis testing as the process of making simultaneous inferences within a model fit to experimental data and reasoning about the implications. To the best of our knowledge, ours is the first work that investigates testing from this perspective, and the first to reach the conclusions that we reach herein.

This paper is organized as follows: We first set the stage for globally-informed evaluation with reusable test collections in Section 2, presenting the idea of statistical hypothesis testing as performing inference in a model fit to evaluation data. In Section 3 we describe the Multiple Comparisons Problem that arises when many hypotheses are tested simultaneously and present formally-motivated ways to address it. These two sections are really just a prelude that summarize widely-known statistical methods, though we expect much of the material to be new to IR researchers. Sections 4 and 5 respectively demonstrate the consequences of these ideas on previous TREC experiments and argue about the implications for the whole evaluation paradigm based on reusable test collections. In Section 6 we conclude with some philosophical musings about statistical analysis in IR.

## 2. STATISTICAL TESTING AS MODEL-BASED INFERENCE

We define a hypothesis test as a procedure that takes data  $y$ ,  $X$  and a *null hypothesis*  $H_0$  and outputs a  $p$ -value, the probability  $P(y|X, H_0)$  of observing the data given that hypothesis. In typical systems-based IR experiments,  $y$  is a matrix of average precision (AP) or other effectiveness evaluation measure values, and  $X$  relates each value  $y_{ij}$  to a particular system  $i$  and topic  $j$ ;  $H_0$  is the hypothesis that knowledge of the system does not inform the value of  $y$ , i.e. that the system has no effect.

Researchers usually learn these tests as a sort of recipe of steps to apply to the data in order to get the  $p$ -value. We are given instructions on how to interpret the  $p$ -value<sup>1</sup>. We are sometimes told some of the assumptions behind the test. We rarely learn the origins of them and what happens when they are violated.

We approach hypothesis testing from a modeling perspective. IR researchers are familiar with modeling: indexing and retrieval algorithms are based on a model of the relevance of a document to a query that includes such features as term frequencies, collection/document frequencies, and document lengths. Similarly, statistical hypothesis tests are implicitly based on models of  $y$  as a function of features derived from  $X$ . Running a test can be seen as fitting a model, then performing inference about model parameters; the recipes we learn are just shortcuts allowing us to perform the inference without fitting the full model. Understanding the models is key to developing a philosophy of IR experimentation based on test collections.

Our goal in this section is to explicate the models behind the t-test and other tests popular in IR. As it turns out, the model implicit in the t-test is a special case of a model we are all familiar with: the linear regression model.

### 2.1 Linear models

Linear models are a broad class of models in which a dependent or response variable  $y$  is modeled as a linear combination of one or more independent variables  $X$  (also called covariates, features, or predictors) [Monahan 2008]. They include models of non-linear transformations of  $y$ , models in which some of the covariates represent “fixed effects” and others represent “random effects”, and models with correlations between covariates or families of covariates [McCullagh and Nelder 1989; Raudenbush and Bryk 2002; Gelman et al. 2004; Venables and Ripley 2002]. It is a very

<sup>1</sup>Though the standard pedagogy on  $p$ -value interpretation is muddled and inconsistent because it combines aspects of two separate philosophies of statistics [Berger 2003].

well-understood, widely-used class of models.

The best known instance of linear models is the multiple linear regression model [Draper and Smith 1998]:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_j + \epsilon_i$$

Here  $\beta_0$  is a model intercept and  $\beta_j$  is a coefficient on independent variable  $x_j$ ;  $\epsilon_i$  is a random error assumed to be normally distributed with mean 0 and uniform variance  $\sigma^2$  that captures everything about  $y$  that is not captured by the linear combination of  $X$ .

Fitting a linear regression model involves finding estimators of the intercept and coefficients. This is done with *ordinary least squares* (OLS), a simple approach that minimizes the sum of squared  $\epsilon_i$ —which is equivalent to finding the maximum-likelihood estimators of  $\beta$  under the error normality assumption [Wasserman 2003]. The coefficients have standard errors as well; dividing an estimator  $\hat{\beta}_j$  by its standard error  $s_j$  produces a statistic that can be used to test hypotheses about the significance of feature  $x_j$  in predicting  $y$ . As it happens, this statistic has a Student’s  $t$  distribution (asymptotically) [Draper and Smith 1998].

Variables can be categorical or numeric; a categorical variable  $x_j$  with values in the domain  $\{e_1, e_2, \dots, e_q\}$  is converted into  $q-1$  mutually-exclusive binary variables  $x_{jk}$ , with  $x_{jk} = 1$  if  $x_j = e_k$  and 0 otherwise. Each of these binary variables has its own coefficient. In the fitted model, all of the coefficients associated with one categorical variable will have the same estimate of standard error.<sup>2</sup>

**2.1.1 Analysis of variance (ANOVA) and the  $t$ -test.** Suppose we fit a linear regression model with an intercept and two categorical variables, one having domain of size  $m$ , the other having domain of size  $n$ . So as not to confuse with notation, we will use  $\mu$  for the intercept,  $\beta_i$  ( $2 \leq i \leq m$ ) for the binary variables derived from the first categorical variable, and  $\gamma_j$  ( $2 \leq j \leq n$ ) for the binary variables derived from the second categorical variable. If  $\beta_i$  represents a system,  $\gamma_j$  represents a topic, and  $y_{ij}$  is the average precision (AP) of system  $i$  on topic  $j$ , then we are modeling AP as a linear combination of a population effect  $\mu$ , a system effect  $\beta_i$ , a topic effect  $\gamma_j$ , and a random error  $\epsilon_{ij}$ .

Assuming every topic has been run on every system (a fully-nested design), the maximum likelihood estimates of the coefficients are:

$$\begin{aligned}\hat{\beta}_i &= \text{MAP}_i - \text{MAP}_1 \\ \hat{\gamma}_j &= \text{AAP}_j - \text{AAP}_1 \\ \hat{\mu} &= \text{AAP}_1 + \text{MAP}_1 - \frac{1}{nm} \sum_{i,j} AP_{ij}\end{aligned}$$

where  $\text{AAP}_j$  is “average average precision” of topic  $j$  averaged over  $m$  systems.

<sup>2</sup>The reason for using  $q-1$  binary variables rather than  $q$  is that the value of any one is known by looking at all the others, and thus one contributes no new information about  $y$ . Generally the first one is eliminated.

Then the residual error  $\epsilon_{ij}$  is

$$\begin{aligned}\epsilon_{ij} &= y_{ij} - (\text{AAP}_1 + \text{MAP}_1 - \frac{1}{nm} \sum_{i,j} AP_{ij} + \text{MAP}_i - \text{MAP}_1 + \text{AAP}_j - \text{AAP}_1) \\ &= y_{ij} - (\text{MAP}_i + \text{AAP}_j - \frac{1}{nm} \sum_{i,j} AP_{ij})\end{aligned}$$

and

$$\hat{\sigma}^2 = \text{MSE} = \frac{1}{(n-1)(m-1)} \sum_{i,j} \epsilon_{ij}^2$$

The standard errors for  $\beta$  and  $\gamma$  are

$$s_\beta = \sqrt{\frac{2\hat{\sigma}^2}{n}} \quad s_\gamma = \sqrt{\frac{2\hat{\sigma}^2}{m}}$$

We can test hypotheses about systems using the estimator  $\hat{\beta}_i$  and standard error  $s_\beta$ : dividing  $\hat{\beta}_i$  by  $s_\beta$  gives a  $t$ -statistic. If the null hypothesis that system  $i$  has no effect on  $y_{ij}$  is true, then  $\hat{\beta}_i/s_\beta$  has a  $t$  distribution. Thus if the value of that statistic is unlikely to occur in that null distribution, we can reject the hypothesis that system  $i$  has no effect.

This is the ANOVA model. It is a special case of linear regression in which all features are categorical and fully nested.

Now suppose  $m = 2$ , that is, we want to perform a paired comparison of two systems. If we fit the model above, we get:

$$\begin{aligned}\hat{\beta}_2 &= \text{MAP}_2 - \text{MAP}_1 \\ \hat{\sigma}^2 &= \frac{1}{n-1} \sum_{j=1}^n \frac{1}{2} ((y_{1j} - y_{2j}) - (\text{MAP}_1 - \text{MAP}_2))^2 \\ s_\beta &= \sqrt{\frac{2\hat{\sigma}^2}{n}}\end{aligned}$$

Note that these are equivalent to the estimates we would use for a two-sided paired  $t$ -test of the hypothesis that the MAPs are equal. Taking  $t = \hat{\beta}_2/s_\beta$  gives a  $t$ -statistic that we can use to test that hypothesis, by finding the  $p$ -value—the probability of observing that value in a null  $t$  distribution. The  $t$ -test is therefore a special case of ANOVA, which in turn is a special case of linear regression.

**2.1.2 Assumptions of the  $t$ -test.** Formulating the  $t$ -test as a linear model allows us to see precisely what its assumptions are:

- (1) errors  $\epsilon_{ij}$  are normally distributed with mean 0 and variance  $\sigma^2$  (normality);
- (2) variance  $\sigma^2$  is constant over systems (homoskedasticity);
- (3) effects are additive and linearly related to  $y_{ij}$  (linearity);
- (4) topics are sampled i.i.d. (independence).

Normality, homoskedasticity, and linearity are built into the model. Independence is an assumption needed for using OLS to fit the model.

	model and standard error $s_\beta$ ( $\times 10^2$ )					
$H_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$	$M_6$
INQ601 = INQ602	<b>0.250</b>	0.082	0.095	0.078	0.118	0.031*
INQ601 = INQ603	0.024*	<b>0.001*</b>	0.001*	0.001*	0.002*	0.000*
INQ601 = INQ604	0.001*	0.000*	<b>0.000*</b>	0.000*	0.000*	0.000*
INQ602 = INQ603	0.248	0.081	0.094	<b>0.077</b>	0.117	0.030*
INQ602 = INQ604	0.031*	0.001*	0.002*	0.001*	<b>0.004*</b>	0.000*
INQ603 = INQ604	0.299	0.117	0.132	0.111	0.158	<b>0.051</b>
$H_0$	$M_7$	$M_8$	$M_9$	$M_{10}$	$M_{11}$	$M_{12}$
INQ601 = INQ602	<b>0.141</b>	<b>0.157</b>	<b>0.069</b>	0.075	<b>0.109</b>	<b>0.479</b>
INQ601 = INQ603	<b>0.004*</b>	0.004*	0.000*	0.001*	<b>0.002*</b>	<b>0.159</b>
INQ601 = INQ604	0.000*	<b>0.000*</b>	<b>0.000*</b>	0.000*	<b>0.000*</b>	<b>0.044*</b>
INQ602 = INQ603	<b>0.140</b>	0.141	0.068	<b>0.071</b>	<b>0.108</b>	<b>0.477</b>
INQ602 = INQ604	0.006*	<b>0.008*</b>	0.001*	<b>0.001*</b>	<b>0.003*</b>	<b>0.181</b>
INQ603 = INQ604	0.184	0.186	<b>0.097</b>	<b>0.105</b>	<b>0.149</b>	<b>0.524</b>

Table I.  $p$ -values for each of the six pairwise hypotheses about TREC-8 UMass systems within each of 11 models fit to different subsets of the four systems, plus a twelfth fit to all 129 TREC-8 systems. Boldface indicates that both systems in the corresponding hypothesis were used to fit the corresponding model. \* denotes significance at the 0.05 level. Depending on the systems used to fit the model, standard error can vary from 0.00749 to 0.02333, and as a result  $p$ -values for pairwise comparisons can vary substantially.

Homoskedasticity and linearity are not true in most IR experiments. This is a simple consequence of effectiveness evaluation measures having a discrete value in a bounded range (nearly always  $[0, 1]$ ). We will demonstrate that they are not true—and the effect of violating them—in depth in Section 4.4 below.

## 2.2 Model-based inference

Above we showed that a  $t$ -test is equivalent to an inference about coefficient  $\beta_2$  in a model fit to evaluation results from two systems over  $n$  topics. In general, if we have a linear model with variance  $\hat{\sigma}^2$  and system standard error  $s_\beta$ , we can perform inference about a difference between any two systems  $i, j$  by dividing  $\text{MAP}_i - \text{MAP}_j$  by  $s_\beta$ . This produces a  $t$  statistic that we can compare to a null  $t$  distribution.

If we can test any hypothesis in any model, what happens when we test one hypothesis under different models? Can we expect to see the same results? To make this concrete, we look at four systems submitted to the TREC-8 ad hoc track<sup>3</sup>: INQ601, INQ602, INQ603, and INQ604, all from UMass Amherst. We can fit a model to any subset of these systems; the system effect estimators  $\hat{\beta}_i$  will be congruent across models. The population and topic effect estimators  $\hat{\mu}_i, \hat{\gamma}_j$  will vary between models depending on the effectiveness of the systems used to fit the model. Therefore the standard error  $s_\beta$  will vary, and it follows that  $p$ -values will vary too.

Table I shows the standard error  $s_\beta$  (times  $10^2$  for clarity) for each model and  $p$ -values for all six pairwise hypotheses under each model. The first six models are fit to just two systems, the next four are fit to three systems, the 11th is fit to all four, and the 12th is fit to all 129 TREC-8 systems. Reading across rows gives an

<sup>3</sup>The TREC-8 data is described in more detail in Section 4.1.

idea of how much  $p$ -values can vary depending on which systems are used to fit the model; for the first hypothesis the  $p$ -values range from 0.031 to 0.250 in the UMass-only model, or 0.479 in the TREC-8 model. For an individual researcher, this could be a difference between attempting to publish the results and dropping a line of research entirely.

Of course, it would be very strange indeed to test a hypothesis about INQ601 and INQ602 under a model fit to INQ603 and INQ604; the key point is that if hypothesis testing is model-based inference, then hypotheses can be tested in *any* model. Instead of using a model fit to just two systems, we should construct a model that best reflects everything we know about the world. The two systems INQ603 and INQ604 can only produce a model with a very narrow view of the world that ignores a great deal about what we know, and that leads to finding many differences between the UMass systems. The full TREC-8 model takes much more information into account, but leads to inferring almost no differences between the same systems.

### 2.3 Other tests, other models

Given the variation in t-test  $p$ -values, then, perhaps it would make more sense to use a different test altogether. It should be noted that in the model-based view, *every* hypothesis test is based on a model, and testing a hypothesis is always equivalent to fitting a model and making inferences from model parameters. The differences between specific tests are mainly in how  $y$  is transformed and in the distributional assumptions they make about the errors.

**2.3.1 Non-parametric tests.** A *non-parametric* test is one in which the error distribution has no parameters that need to be estimated from data [Wasserman 2006]. This is generally achieved by transforming the measurements in some way. A very common transformation is to use ranks rather than the original values; sums of ranks have parameter-free distributional guarantees that sums of values do not have.

The two most common non-parametric tests in IR are the sign test and Wilcoxon's signed rank test. The sign test transforms the data into signs of differences in AP  $\text{sgn}(y_{2j} - y_{1j})$  for each topic. The sign is then modeled as a sum of a parameter  $\mu$  and random error  $\epsilon$ . The transformation into signs gives  $\epsilon$  a binomial distribution with  $n$  trials and success probability  $1/2$  (and recentered to have mean 0).

$$\begin{aligned}\text{sgn}(y_{2j} - y_{1j}) &= \frac{1}{n} (\mu + \epsilon_j) \\ \epsilon_j + n/2 &\sim \text{Binom}(n, 1/2)\end{aligned}$$

The maximum likelihood estimator for  $\mu$  is  $S$ , the number of differences in AP that are positive. To perform inference, we calculate the probability of observing  $S$  in a binomial distribution with parameters  $n, 1/2$ . There is no clear way to generalize this to a model fit to  $m$  systems (the way an ANOVA is a generalization of the t-test), but we can still test hypotheses about any two systems within a model fit to a particular pair.

Wilcoxon's signed rank test transforms unsigned differences in AP  $|y_{2j} - y_{1j}|$  to ranks, then multiplies the rank by the sign. This transformed quantity is modeled

as a sum of a population effect  $\mu$  and random error  $\epsilon$ , which has a symmetric distribution with mean  $n(n+1)/4$  and variance  $n(n+1)(2n+1)/24$  (we will call it the Wilcoxon distribution).

$$\begin{aligned} \text{sgn}(y_{2j} - y_{1j}) \cdot \text{rnk}(|y_{2j} - y_{1j}|) &= \mu - \epsilon_j \\ \epsilon_j &\sim \text{Wilcoxon}(n) \end{aligned}$$

The maximum likelihood estimator for  $\mu$  is the sum of the positively-signed ranks. This again is difficult to generalize to  $m$  systems, though there are other rank-based non-parametric tests for those cases (e.g. Friedman, Mann-Whitney).

*2.3.2 Empirical error distributions.* Instead of transforming  $y$  to guarantee a parameter-free error distribution, we could estimate the error distribution directly from data. Two widely-used approaches to this are Fisher’s permutation procedure and the bootstrap.

Fisher’s permutation procedure produces an error distribution by permuting the assignment of within-topic  $y$  values to systems. This ensures that the system identifier is unrelated to error; we can then reject hypotheses about systems if we find a test statistic computed over those systems in the tail of the error distribution. The permutation distribution can be understood as an  $n \times m \times nm!$  array: for each permutation (of which there are  $nm!$  total), there is an  $n \times m$  table of topic-wise permuted values minus column means. Since there is such a large number of permutations, we usually estimate the distribution by random sampling.

The model is still additive; the difference from the linear model is that no assumptions need be made about the error distribution.

$$\begin{aligned} y_{ij} &= \mu + \beta_i + \gamma_j + \epsilon_{ij} \\ \epsilon_{ij} &\sim \text{Perm}(y, X) \end{aligned}$$

The maximum likelihood estimator of  $\beta_i$  is  $\text{MAP}_i$ . To test a hypothesis about the difference between systems, we look at the probability of observing  $\hat{\beta}_i - \hat{\beta}_j = \text{MAP}_i - \text{MAP}_j$  in the column means of the permutation distribution.

The permutation test relaxes the t-test’s homoskedasticity assumption to a weaker assumption that systems are *exchangeable*, which intuitively means that their order does not matter.

### 3. PAIRWISE SYSTEM COMPARISONS

The following situation is common in IR experiments: take a baseline system  $S_0$  based on a simple retrieval model, then compare it to two or more systems  $S_1, S_2, \dots, S_m$  that build on that model. Those systems or some subset of them will often be compared amongst each other. All comparisons use the same corpora, same topics, and same relevance judgments. The comparisons are evaluated by a paired test of significance such as the t-test, so at least two separate t-tests are performed:  $S_0$  against  $S_1$  and  $S_0$  against  $S_2$ . Frequently the alternatives will be tested against one another as well, so with  $m$  systems the total number of tests is  $m(m-1)/2$ . This gives rise to the Multiple Comparisons Problem: as the number of tests increases, so does the number of false positives—apparently significant differences that are not really different.



In the following sections, we first describe the Multiple Comparisons Problem (MCP) and how it affects IR experimental analysis. We then describe the General Linear Hypotheses (GLH) approach to enumerating the hypotheses to be tested, focusing on three common scenarios in IR experimentation: multiple comparisons to a baseline, sequential comparisons, and all-pairs comparisons. We then show how to resolve MCP by adjusting  $p$ -values using GLH.

### 3.1 The Multiple Comparisons Problem

In a usual hypothesis test, we are evaluating a null hypothesis that the difference in mean performance between two systems is zero (in the two-sided case) or signed in some way (in the one-sided case).

$$\begin{array}{ll} H_0 : \text{MAP}_0 = \text{MAP}_1 & H_0 : \text{MAP}_0 > \text{MAP}_1 \\ H_a : \text{MAP}_0 \neq \text{MAP}_1 & H_a : \text{MAP}_0 \leq \text{MAP}_1 \end{array}$$

As we saw above, we make an inference about a model parameter by calculating a probability of observing its value in a null distribution. If that probability is sufficiently low (usually less than 0.05), we can reject the null hypothesis. The rejection level  $\alpha$  is a false positive rate; if the same test is performed many times on different samples, and if the null hypothesis is actually true (and all assumptions of the test hold), then we will incorrectly reject the null hypothesis in  $100 \cdot \alpha\%$  of the tests—with the usual 0.05 level, we will conclude the systems are different when they actually are not in 5% of the experiments.

Since the error rate is a percentage, it follows that the *number* of errors increases as the number of tests performed increases. Suppose we are performing pairwise tests on a set of seven systems, for a total of 21 pairwise tests. If these systems are all equally effective, we should *expect* to see one erroneous significant difference at the 0.05 level. In general, with  $k$  pairwise tests the probability of finding at least one incorrect significant result at the  $\alpha$  significance level increases to  $1 - (1 - \alpha)^k$ . This is called the *family-wise error rate*, and it is defined to be the probability of at least one false positive in  $k$  experiments [Miller 1981]. Figure 1 shows the rate of increase in the family-wise error rate as  $k$  goes from one experiment to a hundred.

This is problematic in IR because portable test collections make it very easy to run many experiments and therefore very likely to find falsely significant results. With only  $m = 12$  systems and  $k = 66$  pairwise tests between them, that probability is over 95%; the expected number of significant results is 3.3. This can be (and often is) mitigated by showing that results are consistent over several different collections, but that does not fully address the problem. We are interested in a formal model of and solution to this problem.

### 3.2 General Linear Hypotheses

We will first formalize the notion of performing multiple tests using the idea of General Linear Hypotheses [Mardia et al. 1980]. The usual paired t-test we describe above compares two means within a model fit to just those two systems. The more general ANOVA fit to  $m$  systems tests the so-called *omnibus hypothesis* that all MAPs are equal:

$$H_0 : \text{MAP}_0 = \text{MAP}_1 = \text{MAP}_2 = \dots = \text{MAP}_{m-1}$$

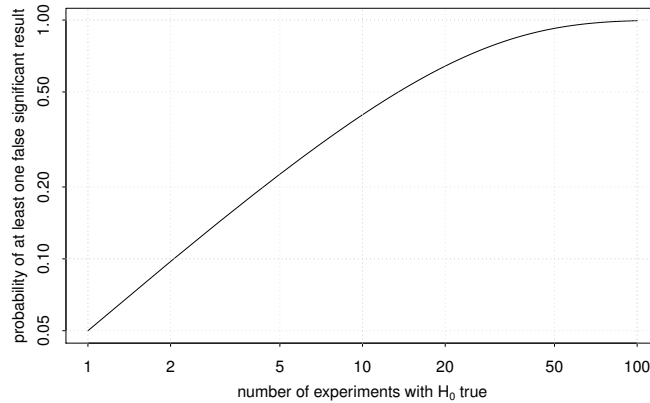


Fig. 1. The probability of falsely rejecting  $H_0$  (with  $p < 0.05$ ) at least once increases rapidly as the number of experiments for which  $H_0$  is true increases.

The alternative is that some pair of MAPs are not equal. We want something in between: a test of two or more pairwise hypotheses within a model that has been fit to all  $m$  systems.

As we showed above, the coefficients in the linear model that correspond to systems are exactly equal to the difference in means between each system and the baseline (or first) system in the set. We will denote the vector of coefficients  $\beta$ ; it has  $m$  elements, but the first is the model intercept rather than a coefficient for the baseline system. We can test hypotheses about differences between each system and the baseline by looking at the corresponding coefficient and its  $t$  statistic. We can test hypotheses about differences between arbitrary pairs of systems by looking at the difference between the corresponding coefficients; since they all have the same standard error  $s_\beta$ , we can obtain a  $t$  statistic by dividing any difference by  $s_\beta$ .

We will formalize this with matrix multiplication. Define a  $k \times m$  *contrast matrix*  $\mathbf{K}$ ; each column of this matrix corresponds to a system (except for the first column, which corresponds to the model intercept), and each row corresponds to a hypothesis we are interested in. Each row will have a 1 and a  $-1$  in the cells corresponding to two systems to be compared to each other, or just a 1 in the cell corresponding to a system to compare to the baseline. If  $\mathbf{K}$  is properly defined, the matrix-vector product  $\frac{1}{s_\beta} \mathbf{K} \beta$  produces a vector of  $t$  statistics for our hypotheses.

**3.2.1 Multiple comparisons to a baseline system.** In this scenario, a researcher has a baseline  $S_0$  and compares it to one or more alternatives  $S_1, \dots, S_{m-1}$  with paired tests. The correct analysis would first fit a model to data from all systems, then compare the system coefficients  $\beta_1, \dots, \beta_{m-1}$  (with standard error  $s_\beta$ ). As discussed above, since we are performing  $m$  tests, the family-wise false positive rate increases to  $1 - (1 - \alpha)^m$ .

The contrast matrix  $\mathbf{K}$  for this scenario looks like this:

$$\mathbf{K}_{\text{base}} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdots & & & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Multiplying  $\mathbf{K}_{\text{base}}$  by  $\beta$  produces the vector  $[\beta_1 \ \beta_2 \ \cdots \ \beta_{m-1}]'$ , and then dividing by  $s_\beta$  produces the  $t$  statistics.

**3.2.2 Sequential system comparisons.** In this scenario, a researcher has a baseline  $S_0$  and sequentially develops alternatives, testing each one against the one that came before, so the sequence of null hypotheses is:

$$S_0 = S_1; S_1 = S_2; \cdots; S_{m-2} = S_{m-1}$$

Again, since this involves  $m$  tests, the family-wise false positive rate is  $1 - (1 - \alpha)^m$ .

For this scenario, the contrast matrix is:

$$\mathbf{K}_{\text{seq}} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \cdots & & & & & & \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

and

$$\frac{1}{s_\beta} \mathbf{K}_{\text{seq}} \beta = \left[ \frac{\beta_1}{s_\beta} \quad \frac{\beta_2 - \beta_1}{s_\beta} \quad \frac{\beta_3 - \beta_2}{s_\beta} \quad \cdots \quad \frac{\beta_{m-1} - \beta_{m-2}}{s_\beta} \right]'$$

which is the vector of  $t$  statistics for this set of hypotheses.

Incidentally, using significance as a condition for stopping sequential development leads to the *sequential testing problem* [Wald 1947], which is related to MCP but outside the scope of this work.

**3.2.3 All-pairs system comparisons.** Perhaps the most common scenario is that there are  $m$  different systems and the researcher performs all  $m(m-1)/2$  pairwise tests. The family-wise false positive rate is  $1 - (1 - \alpha)^{m(m-1)/2}$ , which as we suggested above rapidly approaches 1 with increasing  $m$ .

Again, we assume we have a model fit to all  $m$  systems, providing coefficients  $\beta_1, \dots, \beta_{m-1}$  and standard error  $s_\beta$ . The contrast matrix will have a row for every

pair of systems, including the baseline comparisons in  $\mathbf{K}_{\text{base}}$ :

$$\mathbf{K}_{\text{all}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdots & & & & & & & \\ 0 & -1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & \cdots & 0 & 0 \\ \cdots & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

with

$$\frac{1}{s_\beta} \mathbf{K}_{\text{all}} \beta = \left[ \frac{\beta_1}{s_\beta} \cdots \frac{\beta_{m-1}}{s_\beta} \frac{\beta_2 - \beta_1}{s_\beta} \frac{\beta_3 - \beta_1}{s_\beta} \cdots \frac{\beta_{m-1} - \beta_1}{s_\beta} \cdots \frac{\beta_{m-1} - \beta_{m-2}}{s_\beta} \right]'$$

giving the  $t$  statistics for all  $m(m-1)/2$  hypotheses.

### 3.3 Adjusting $p$ -values for multiple comparisons

GLH enumerates  $t$  statistics for all of our hypotheses, but if we evaluate their probabilities naïvely, we run into MCP. The idea of  $p$ -value adjustment is to modify the  $p$ -values so that the false positive rate is 0.05 for the entire *family* of comparisons, i.e. so that there is only a 5% chance of having at least one falsely significant result in the full set of tests. We present two approaches, one designed for all-pairs testing with  $\mathbf{K}_{\text{all}}$  and one for arbitrary  $\mathbf{K}$ . Many approaches exist, from the very simple and very weak Bonferroni correction [Miller 1981] to very complex, powerful sampling methods; the ones we present are relatively simple while retaining good statistical power.

**3.3.1 Tukey's Honest Significant Differences (HSD).** John Tukey developed a method for adjusting  $p$ -values for all-pairs testing based on a simple observation: within the linear model, if the largest mean difference is not significant, then none of the other mean differences should be significant either [Hsu 1996]. Therefore all we need to do is formulate a null distribution for the largest mean difference. Any observed mean difference that falls in the tail of that distribution is significant. The family-wise error rate will be no greater than 0.05, because the maximum mean difference will have an error rate of 0.05 by construct and errors in the maximum mean difference dominate all other errors.

At first this seems simple: a null distribution for the largest mean difference could be the same as the null distribution for any mean difference, which is a  $t$  distribution centered at 0 with  $n-1$  degrees of freedom. But it is actually *unlikely* that the largest mean difference would be zero, even if the omnibus null hypothesis is true. The expected maximum mean difference is a function of  $m$ , increasing as the number of systems increases. Figure 2 shows an example distribution with

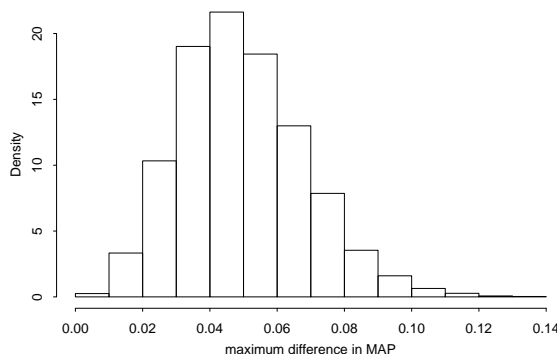


Fig. 2. Empirical studentized range distribution of maximum difference in MAP among five systems under the omnibus null hypothesis that there is no difference between any of the five systems. Difference in MAP would have to exceed 0.08 to reject any pairwise difference between systems.

$m = 5$  systems over 50 topics.<sup>4</sup>

The maximum mean difference is called the *range* of the sample and is denoted  $r$ . Just as we divide mean differences by  $s_\beta = \sqrt{2\hat{\sigma}^2/n}$  to obtain a *studentized mean*  $t$ , we can divide  $r$  by  $\sqrt{\hat{\sigma}^2/n}$  to obtain a *studentized range*  $q$ . The distribution of  $q$  is called the *studentized range distribution*; when comparing  $m$  systems over  $n$  topics, the studentized range distribution has  $m, (n - 1)(m - 1)$  degrees of freedom.

Once we have a studentized range distribution based on  $r$  and the null hypothesis, we can obtain a  $p$ -value for each pairwise comparison of systems  $i, j$  by finding the probability of observing a studentized mean  $t_{ij} = (\hat{\beta}_i - \hat{\beta}_j)/\sqrt{\hat{\sigma}^2/n}$  given that distribution. The range distribution has no closed form that we are aware of, but distribution tables are available, as are cumulative probability and quantile functions in most statistical packages.

Thus computing the Tukey HSD  $p$ -values is largely a matter of trading the usual  $t$  distribution with  $n - 1$  degrees of freedom for a studentized range distribution with  $m, (n - 1)(m - 1)$  degrees of freedom. The  $p$ -values obtained from the range distribution will never be less than those obtained from a  $t$  distribution. When  $m = 2$  they will be equivalent.

**3.3.2 Single-step method using a multivariate  $t$  distribution.** A more powerful approach to adjusting  $p$ -values uses information about correlations between hypotheses [Hothorn et al. 2008; Bretz et al. 2010]. If one hypothesis involves systems  $S_1$  and  $S_2$ , and another involves systems  $S_2$  and  $S_3$ , those two hypotheses are correlated; we expect that their respective  $t$  statistics would be correlated as well. When computing the probability of observing a  $t$ , we should take this correlation into account. A multivariate generalization to the  $t$  distribution allows us to do so.

<sup>4</sup>We generated the data by repeatedly sampling 5 APs uniformly at random from all 129 submitted TREC-8 runs for each of the 50 TREC-8 topics, then calculating MAPs and taking the difference between the maximum and the minimum values. Our sampling process guarantees that the null hypothesis is true, yet the expected maximum difference in MAP is 0.05.

Let  $\mathbf{K}'$  be identical to a contrast matrix  $\mathbf{K}$ , except that the first column represents the baseline system (so any comparisons to the baseline have a  $-1$  in the first column). Then take  $\mathbf{R}$  to be the correlation matrix of  $\mathbf{K}'$  (i.e.  $\mathbf{R} = \mathbf{K}'\mathbf{K}'^T/(k-1)$ ). The correlation between two identical hypotheses is 1, and the correlation between two hypotheses with one system in common is 0.5 (which will be negative if the common system is negative in one hypothesis and positive in the other).

Given  $\mathbf{R}$  and a  $t$  statistic  $t_i$  from the GLH above, we can compute a  $p$ -value as  $1 - P(T < |t_i| \mid \mathbf{R}, \nu)$ , where  $\nu$  is the number of degrees of freedom  $\nu = (n-1)(m-1)$  and  $P(T < |t_i|)$  is the cumulative density of the multivariate  $t$  distribution:

$$P(T < |t_i| \mid \mathbf{R}, \nu) = \int_{-t_i}^{t_i} \int_{-t_i}^{t_i} \cdots \int_{-t_i}^{t_i} P(x_1, \dots, x_k \mid \mathbf{R}, \nu) dx_1 \dots dx_k$$

Here  $P(x_1, \dots, x_k \mid \mathbf{R}, \nu) = P(\mathbf{x} \mid \mathbf{R}, \nu)$  is the  $k$ -variate  $t$  distribution with density function

$$P(\mathbf{x} \mid \mathbf{R}, \nu) \propto (\nu\pi)^{-k/2} |\mathbf{R}|^{-1/2} \left(1 + \frac{1}{\nu} \mathbf{x}^T \mathbf{R}^{-1} \mathbf{x}\right)^{-(\nu+k)/2}$$

Unfortunately the integral can only be computed by numerical methods that grow computationally-intensive and more inaccurate as the number of hypotheses  $k$  grows. This method is only useful for relatively small  $k$ , but it is more powerful than Tukey's HSD because it does not assume that we are interested in comparing all pairs, which is the worst case scenario for MCP. We have used it to test up to  $k \approx 200$  simultaneous hypotheses in a reasonable amount of time (an hour or so); Appendix B shows how to use it in the statistical programming environment R.

#### 4. APPLICATION TO TREC

We have now seen that inference about differences between systems can be affected both by the model used to do the inference and by the number of comparisons that are being made. The theoretical discussion raises the question of how this would affect the analysis of large experimental settings like those done at TREC. What happens if we fit models to full sets of TREC systems rather than pairs of systems? What happens when we adjust  $p$ -values for thousands of experiments rather than just a few? In this section we explore these issues.

##### 4.1 TREC data

TREC experiments typically proceed as follows: organizers assemble a corpus of documents and develop information needs for a search task in that corpus. Topics are sent to research groups at sites (universities, companies) that have elected to participate in the track. Participating groups run the topics on one or more retrieval systems and send the retrieved results back to TREC organizers. The organizers then obtain relevance judgments that will be used to evaluate the systems and test hypotheses about retrieval effectiveness.

The data available for analysis, then, is the document corpus, the topics, the relevance judgments, and the retrieved results submitted by participating groups. For this work we use data from the TREC-8 ad hoc task: a newswire collection of around 500,000 documents, 50 topics (numbered 401-450), 86,830 total relevance judgments for the 50 topics, and 129 runs submitted by 40 groups [Voorhees and

Harman 1999]. With 129 runs, there are up to 8,256 paired comparisons. Each group submitted one to five runs, so within a group there are between zero and 10 paired comparisons. The total number of within-group paired comparisons (summing over all groups) is 190.

We used TREC-8 because it is the largest (in terms of number of runs), most extensively-judged collection available. Our conclusions apply to any similar experimental setting.

## 4.2 Analyzing TREC data

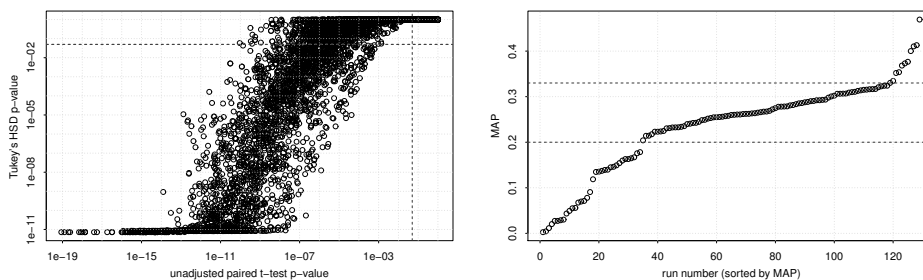
There are several approaches we can take to analyzing TREC data depending on which systems we use to fit a model and how we adjust  $p$ -values from inferences in that model:

- (1) Fit a model to all systems, and then
  - (a) adjust  $p$ -values for all pairwise comparisons in this model (the “honest” way to do all-pairs comparisons between all systems); or
  - (b) adjust  $p$ -values for all pairwise comparisons within participating groups (considering each group to have its own set of experiments that are evaluated in the context of the full TREC experiment); or
  - (c) adjust  $p$ -values independently for all pairwise comparisons within participating groups (considering each group to have its own set of experiments that are evaluated independently of one another).
- (2) Fit separate models to systems submitted by each participating group and then
  - (a) adjust  $p$ -values for all pairwise comparisons within group (pretending that each group is “honestly” evaluating its own experiences out of the context of TREC).

Additionally, we may choose to exclude outlier systems from the model-fitting stage, or fit the model only to the automatic systems or some other subset of systems. For this work we elected to keep all systems in.

Option (2a) is probably the least “honest”, since it excludes so much of the data for fitting the model, and therefore ignores so much information about the world. It is, however, the approach that groups reusing a collection post-TREC would have to take—we discuss this in Section 5.2 below. The other options differ depending on how we define a “family” for the family-wise error rate. If the family is all experiments within TREC (Option (1a)), then the  $p$ -values must be adjusted for  $\binom{m}{2}$  comparisons, which is going to be a very harsh adjustment. If the family is all within-group comparisons (Option (1b)), there are only  $\sum \binom{m_i}{2}$  comparisons, where  $m_i$  is the number of systems submitted by group  $i$ , but the reference becomes the largest difference within any group—so that each group is effectively comparing its experiments against whichever group submitted two systems with the biggest difference in effectiveness. Defining the family to be the comparisons within one group (Option (1c)) may be most appropriate for allowing groups to test their own hypotheses without being affected by what other groups did.

We will investigate all four of the approaches above, comparing the adjusted  $p$ -values to those obtained from independent paired t-tests. We use Tukey’s HSD for Option (1a), since the single-step method is far too computationally-intensive for



(a) Independent t-test  $p$ -values versus Tukey’s HSD  $p$ -values (on log scales to emphasize the range normally considered significant).

(b) Dashed lines indicating the range that would have to be exceeded for a difference to be considered “honestly” significant.

Fig. 3. Analysis of Option (1a), using Tukey’s HSD to adjust  $p$ -values for 8,256 paired comparisons of TREC-8 runs. Together these show that adjusting for MCP across all of TREC results in a radical reconsideration in what would be considered significant.

that many simultaneous tests. We use single-step adjustment for the others, since Tukey’s HSD  $p$ -values would be the same whether we are comparing all pairs or just some subset of all pairs.

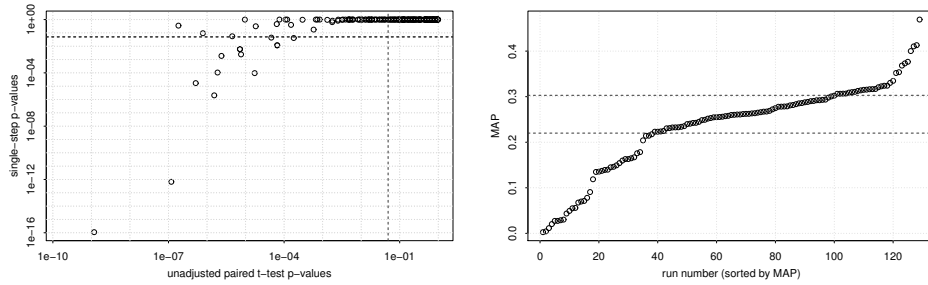
#### 4.3 Analysis of TREC-8 ad hoc experiments

We first analyze Option (1a) of adjusting for all experiments on the same data. Figure 3(a) compares t-test  $p$ -values to Tukey’s HSD  $p$ -values for all 8,256 pairs of TREC-8 submitted runs. The dashed lines are the  $p < 0.05$  cutoffs. Note the HSD  $p$ -values are always many of orders magnitude greater. It would take an absolute difference in MAP greater than 0.13 to find two systems “honestly” different. This is quite a large difference; Figure 3(b) shows the MAPs of all TREC-8 runs with dashed lines giving an idea of a 0.13 difference in MAP. There would be no significant differences between any pair of runs with MAPs from 0.2 to 0.33, a range that includes 65% of the submitted runs. For systems with “average” effectiveness—those with MAPs in the same range—this corresponds to a necessary improvement of 40–65% to find significance! Fully 55% of the significant differences by paired t-tests would be “revoked” by Tukey’s HSD. This is obviously an extreme reconsideration of what it means to be significant.

Next we consider Option (1b) of “global” adjustment of within-group comparisons. Figure 4(a) compares t-test  $p$ -values to single-step  $p$ -values for 190 within-group comparisons (i.e. comparing pairs of runs submitted by the same group; never comparing two runs submitted by two different groups). We still see that many pairwise comparisons that had previously been significant cannot be considered so when adjusting for MCP. It now takes a difference in MAP of about 0.083 to find two systems submitted by the same group significantly different; this corresponds to a 25–40% improvement over an average system. Figure 4(b) illustrates this difference among all 129 TREC-8 runs; 54% of TREC systems fall between the dashed lines. But now 78% of previously-significant comparisons (within groups) would no longer be significant, mainly because groups tend to submit systems that are similar to one another.

Option (1c) uses “local” (within-group) adjustment within a “global” model.

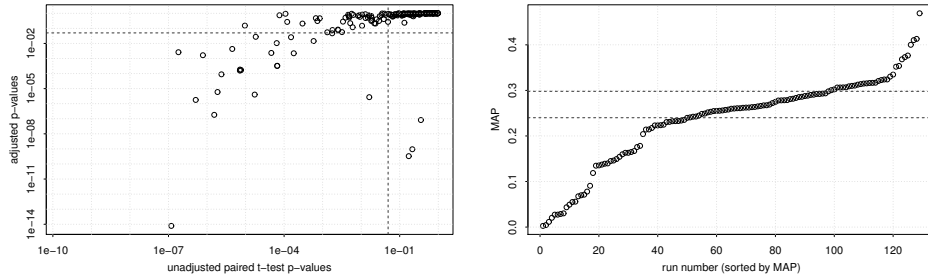




(a) Independent t-test  $p$ -values versus single-step  $p$ -values (on log scales to emphasize the range normally considered significant).

(b) Dashed lines indicating the range that would have to be exceeded for a difference to be considered “honestly” significant.

Fig. 4. Analysis of Option (1b), using “global” single-step  $p$ -value adjustment of 190 within-group comparisons in a model fit to all 129 submitted runs.



(a) Independent t-test  $p$ -values versus single-step  $p$ -values (on log scales to emphasize the range normally considered significant).

(b) Dashed lines indicating the range that would have to be exceeded for a difference to be considered “honestly” significant.

Fig. 5. Analysis of Option (1c), using “local” single-step  $p$ -value adjustment of 190 within-group comparisons in a model fit to all 129 submitted runs.

Figure 5(a) compares t-test  $p$ -values to the single-step adjusted values for the 190 within-group comparisons. It is again the case that many pairwise comparisons that had previously been significant can no longer be considered so, though the effect is now much less harsh: it now takes a difference in MAP of about 0.058 to find two systems submitted by the same group significantly different, which corresponds to a 20–30% improvement over an average system. Figure 5(b) illustrates this difference among the 129 TREC-8 runs; now 37% of TREC systems fall between the dashed lines. There is still a large group of previously-significant comparisons that would no longer be considered significant—66% of the total.

Note that this option presents three cases of pairwise comparisons that are *not* significant with paired t-tests, but that *are* significant in the global model with local adjustment—the three points in the lower right region in Figure 5(a). This is because fitting a model to just those pairs of systems (as the usual t-test does) substantially overestimates the error variance compared to fitting a model to all systems, and thus results in inconclusive evidence to reject the null hypothesis.

Finally we look at Option (2a), using “local” adjustment of “local” (within-

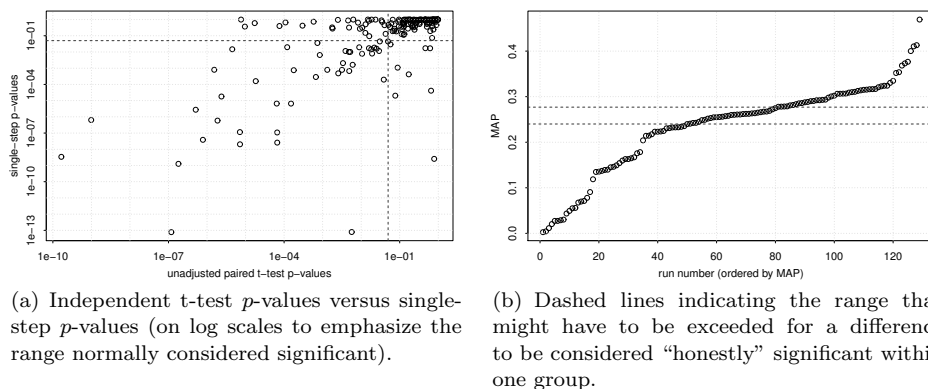


Fig. 6. Analysis of Option (2a), using “local” single-step  $p$ -value adjustment within models fit to each of the 40 groups independently.

group) models. Figure 6(a) compares t-test  $p$ -values to adjusted values for the 190 within-group comparisons. With this analysis we retain many of the significant comparisons that we had with a standard t-test—only 37% are no longer considered significant, and we actually gain eight new significant results for the same reason as given above—but again, this is the least “honest” of our options, since it ignores so much information about topic effects. While the range of MAP required for significance will vary by site, the average is about 0.037, as shown in Figure 6(b). This corresponds to a 10–20% improvement over an average system, which is concomitant with the 10% rule-of-thumb given by Buckley and Voorhees [2000].

It is important to note that once we have decided on one of these options, we cannot go back—if we do, we perform more comparisons, and therefore incur an even higher probability of finding false positives. This is an extremely important point. Once the analysis is done, it must be considered truly *done*, because re-doing the analysis with different decisions is exactly equal to running another batch of comparisons.

#### 4.4 Testing assumptions of the linear model

In Section 2.1.2 we listed the assumptions of the linear model:

- (1) errors  $\epsilon_{ij}$  are normally distributed with mean 0 and variance  $\sigma^2$  (normality);
- (2) variance  $\sigma^2$  is constant over systems (homoskedasticity);
- (3) effects are additive and linearly related to  $y_{ij}$  (linearity);
- (4) topics are sampled i.i.d. (independence).

Do these assumptions hold for TREC data? We do not even have to test them empirically—homoskedasticity definitely does *not* hold, and linearity is highly suspect. Normality probably does not hold either, but not for the reasons usually given [van Rijsbergen 1979]. Independence is unlikely to hold, given the process by which TREC topics are developed, but it is also least essential for the model (being required more for fitting the model than actually modeling the data).

The reason that the first three do not hold in typical IR evaluation settings is actually because our measures are not real numbers in the sense of falling on the real line between  $(-\infty, \infty)$ ; they are discrete numbers falling in the range  $[0, 1]$ . For homoskedasticity, consider a bad system, one that performs poorly on all sampled topics. Since AP on each topic is bounded from below, there is a limit to how bad it can get. As it approaches that lower bound, its variance necessarily decreases; in the limit, if it doesn't retrieve any relevant documents for any topic, its variance is zero.<sup>5</sup> A similar argument holds systems approaching MAP=1, though those are of course rare in reality. The correlation between MAP and variance in AP for TREC-8 runs is over 0.8, supporting this argument.

For linearity the argument may be even more obvious. Anyone with a basic knowledge of linear regression would question the use of it to model bounded values—it is quite likely that it will result in predicted values that fall outside the bounds. Yet we do exactly that every time we use a t-test! In fact, nearly 10% of the fitted values in a full TREC-8 model fall outside the range  $[0, 1]$ . Additivity is less clear in general, but we can point to particular measures that resist modeling as a sum of population, system, and topic effects: recall is dependent on the number of relevant documents in a non-linear, non-additive way. GMAP as well is clearly non-additive, since it is defined as the root of a product of APs.

Normality probably does not hold either, but contrary to previous literature in IR on the t-test, it is not because AP is not normal but simply because a normal distribution is unbounded. In practice, though, sample variances are so low that values outside of the range  $[0, 1]$  are extremely unlikely. Then the Central Limit Theorem (CLT) says that sums of sufficient numbers of independent random variables converge to normal distributions; since  $\epsilon_{ij}$  is the sum of many APs, it follows that  $\epsilon$  is approximately normal. If normal errors are not observed in practice, it is more likely because homoskedasticity and/or independence do not hold.

Violations of these assumptions create an additional source of random variation that affects  $p$ -values, and thus another way that we can incorrectly find a significant result when one does not exist. Thus the interesting question is not whether the assumptions are true or not (for they are not, almost by definition), but whether their violation hurts the performance of the test.

*4.4.1 Consequences of assumption violations.* When investigating violations of test assumptions, there are two issues to consider: the effect on test accuracy and the effect on test power. We focus on accuracy because it is more straightforward to reason about and more directly the focus of this work, but we do not discount the importance of power.

Accuracy is directly related to false positive rate: a less accurate test rejects the null hypothesis when it is true more often than a more accurate test. Ideally accuracy rate is equal to  $1-\alpha$ , the expected false positive rate. To evaluate accuracy, we can randomly generate data in such a way that the omnibus hypothesis is true (i.e. all system means are equal), then fit a model and test that hypothesis. Over

---

<sup>5</sup>One might argue that the worst case is that it retrieves all relevant documents at the bottom of the ranked list. But this is even worse for homoskedasticity—in that case variance is actually a function of relevant document counts rather than anything about the system.

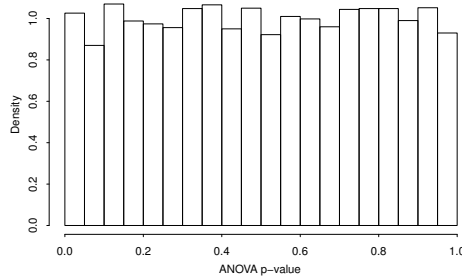
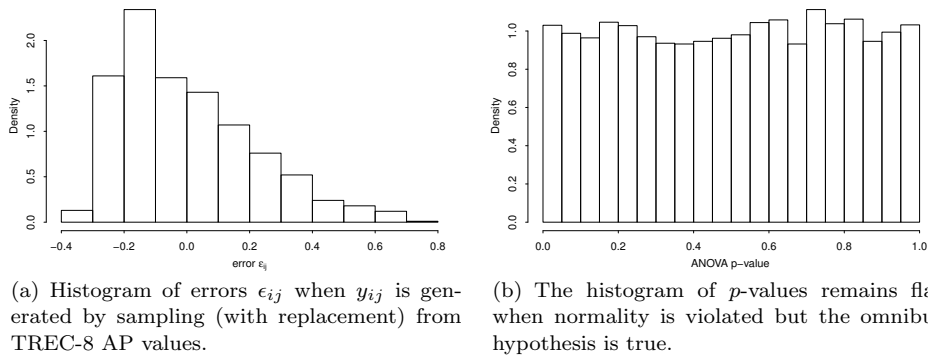


Fig. 7. A uniform distribution of  $p$ -values occurs when assumptions are perfectly satisfied. A Kolmogorov-Smirnoff goodness-of-fit test cannot reject the hypothesis that this is uniform.



(a) Histogram of errors  $\epsilon_{ij}$  when  $y_{ij}$  is generated by sampling (with replacement) from TREC-8 AP values.

(b) The histogram of  $p$ -values remains flat when normality is violated but the omnibus hypothesis is true.

Fig. 8. Analysis of the effect of violations of error normality.

many trials, we obtain a distribution of  $p$ -values. That distribution is expected to be uniform; if it is not, some assumption has been violated.

We start by illustrating the  $p$ -value distribution with data that perfectly meets the assumptions. We sample values  $y_{ij}$  from a normal distribution with mean 0.23 and standard deviation 0.22 (the mean and variance of AP over all TREC-8 submissions). All assumptions are met: since system and topic effects are 0 in expectation, the error distribution is just the sampling distribution recentered to 0; variance is constant by construct; there are no effects for which we have to worry about linearity; and all values are sampled i.i.d. The distribution of  $p$ -values over 10,000 trials is shown in Figure 7. It is flat; a Kolmogorov-Smirnoff test for goodness-of-fit cannot reject the hypothesis that it is uniform.

We relax the normality assumption by sampling from actual AP values rather than a normal distribution: we simply repeatedly sample with replacement from all TREC-8 AP values, then find the ANOVA  $p$ -value. We do not distinguish samples by system or topic, so the error distribution will be the same as the bootstrap AP distribution (recentered to 0). This is shown in Figure 8(a). The resulting  $p$ -value distribution is shown in Figure 8(b). It is still flat, and we still cannot detect any deviation from uniformity, suggesting that the test is robust to the violation of that assumption that we have in IR.

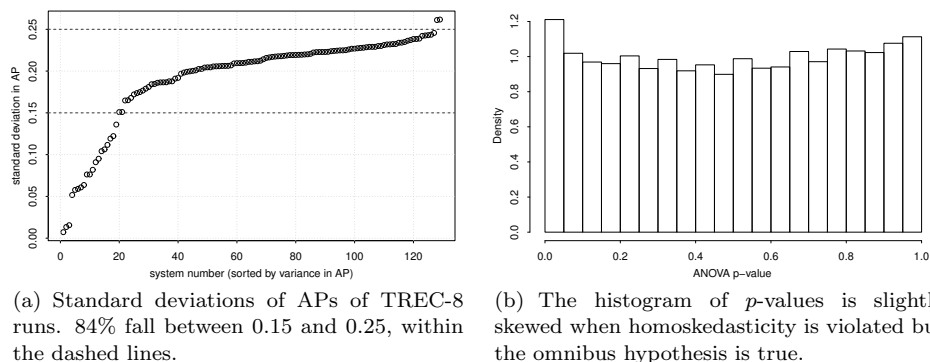


Fig. 9. Analysis of the effect of violations of homoskedasticity.

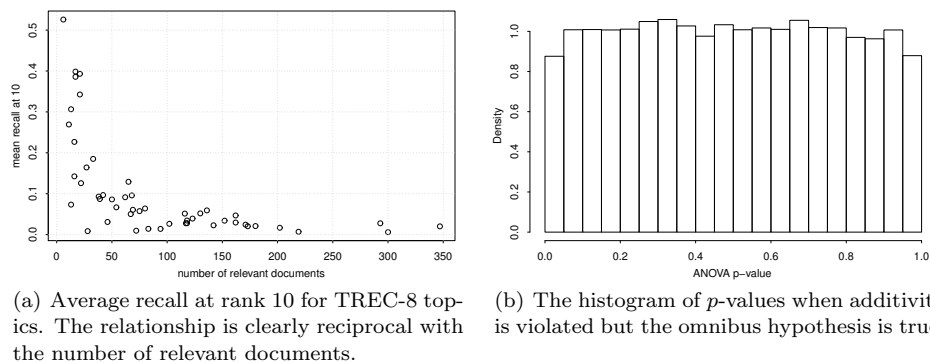


Fig. 10. Analysis of the effect of violations of additivity of effects.

Next we relax the constant variance assumption. TREC-8 systems’ standard deviations range from 0.007 up to 0.262, with 84% falling between 0.15 and 0.25 (Figure 9(a)). To simulate this, we sample data from  $m$  beta distributions with parameters increasing so that their means remain the same but their variances decrease. The resulting  $p$ -value distribution is not quite uniform by inspection (Figure 9(b)), and uniformity is rejected by the Kolmogorov-Smirnoff test. The false error rate is only about 6% compared to the expected 5%, which works out to about one extra false positive in every 100 experiments. Thus a violation of heteroskedasticity does have an effect, but a fairly small one. The effect becomes more pronounced as we sample variances from a wider range.

Finally we relax linearity by sampling values such that they are reciprocally related to topic number (by sampling from recall at rank 10, which has a reciprocal relationship with the number of relevant documents (Figure 10(a))). Interestingly, this causes  $p$ -values to concentrate slightly more around the center (Figure 10(b)); uniformity is again rejected by the Kolmogorov-Smirnoff test. The false error rate drops to about 4.4%. While this technically makes the test “more accurate”, it also reduces its power to detect truly significant differences.

These simulation experiments suggest that if we are going to worry about as-

sumptions, homoskedasticity and linearity rather than normality are the ones we should worry about, though even then the errors are small and may very well cancel out. If we are worried about falsely detecting significant results, MCP is a far greater threat than any assumption made by the linear model.

## 5. DISCUSSION

Our results above raise many questions about the nature of statistical analysis and its effect on the actual research people decide to do.

### 5.1 Significance and publication

As we illustrated in Section 4, the biggest consequence of fitting models and adjusting  $p$ -values for a large set of experiments is that  $p$ -values will decrease dramatically. If everyone were to use the analysis we propose, the biggest consequence would be far fewer significant results. Suppose (hypothetically) that program committees and journal editors use lack of significance as a sufficient condition for rejection, i.e. papers that do not show significant improvements over a baseline are rejected, while those that do show significant improvements may or may not be rejected depending on other factors. Then imagine an alternate universe in which everyone doing research in IR for the last 10 years has been using the methods we present.

First, every paper that was rejected (or not even submitted) because of lack of significance in our universe is still unpublished in this alternate universe. But many of the papers that were *accepted* in our universe have never been published in the alternate universe. The question, then, is: what is the net value of the papers that were accepted in our universe but rejected in the alternate universe? If we knew the answer, we would begin to get a sense of the cost of statistical significance failing to identify an interesting result.

On the other hand, how much time have IR researchers in “our” universe spent on determining that work published due primarily to a small but apparently-significant improvement is not actually worthwhile? How much time reading, reviewing, and re-implementing would be saved if those papers had not been submitted because the authors knew they had no chance to be accepted? This begins to give a sense of the cost of falsely identifying an interesting result with improper statistical analysis.

Finally, if we suppose that re-implementing other published work is the primary time and cost sink, how much time do researchers in the alternate universe spend re-implementing ideas that others have tried but not published due to lack of significance, and how does it compare to the time spent by researchers in our universe re-implementing ideas that were published with reported significant improvements?

We do not claim to know the answers to these questions, but they are important if we are to determine how statistical significance should be taken into account in publication decisions.

### 5.2 Families of experiments and family-wise error rates

How exactly do we define a “family” of experiments for a family-wise error rate? We presented some possibilities in Section 3 with contrast matrices  $\mathbf{K}_{\text{base}}$ ,  $\mathbf{K}_{\text{seq}}$ ,  $\mathbf{K}_{\text{all}}$ . But choosing between those is a minor detail compared to the question of the *breadth* of a family, essentially the choice of the options we considered in Section 4.2.

Let us list some possibilities. A family could be:

- (1) a small set of experiments for one high-level but specific hypothesis conducted by one individual or small research group, for example “smoothing a language model with document clusters can help ad hoc retrieval”;
- (2) a larger set of experiments for a higher-level hypothesis conducted by many individuals and research groups, for example “document clustering can help ad hoc retrieval”;
- (3) a very large set of experiments for many different hypotheses conducted by many research groups for the same broad task, for example “there are techniques that can help ad hoc retrieval”;
- (4) all experiments for all hypotheses that have ever been formulated.

Again, we do not know what the correct formulation is. What we do know is that as we increase the breadth of a family, we decrease the number of significant results that will be found. It could be the case that defining “family” at the level of TREC will completely destroy the ability to detect significance for specific hypotheses about things such as clustering. Would this be useful to the progress of the field? We suspect not, but it is hard to be sure.

Armstrong et al. [2009b] seem to favor Option (3). Achieving that will require a public repository where groups can (and must) upload their runs. These runs would be used to fit a large model within which new hypotheses would be tested. A site such as <http://evaluatir.org> [Armstrong et al. 2009a] could be used for this purpose.

Today, it only makes pragmatic sense to define a family as the first option, since individuals or research groups cannot be aware of all other experiments being conducted around the world. That option essentially has the scope of no more than a handful of publications. Yet researchers do not currently even make corrections at that level. We demonstrate how to do so in Appendix B below.

### 5.3 “Learning” from reusable test collections

We are all aware of the need for separate training, validation, and testing sets of data for developing retrieval models: without these splits, our algorithms will overfit and therefore generalize poorly. The same applies to evaluation. When we evaluate against the same test collections repeatedly, we as individuals and as a community “learn” what works and what doesn’t on those collections; over time we should expect our methods to become more tuned to those collections and less generalizable to other collections. This will manifest as more significant differences over time independently of those predicted by MCP.

The hypothesis is this: due to learning effects, the probability of finding at least one false positive increases over time. If we model these effects, the probability of finding at least one false positive increases at a slower rate.

It is hard to model a learning effect, much less find evidence of it in published literature—we considered mining IR venues for papers that reported significant improvements on TREC datasets, but determining whether there is an effect would require models of the frequency of use of those datasets (we assume usage declines over time, but at what rate?), the rate of increase in IR research over time (it has been increasing, which means more people are using those data sets), and myriad other factors that would influence the rate at which significant improvements are

reported. This is out of the scope of this work, but we mention it as yet another effect on  $p$ -values that we should understand when interpreting them.

#### 5.4 Alternatives to the linear model

What we have presented here is an improvement over fitting  $O(m^2)$  linear models independently, but it is certainly not the best-fitting model. If the linear model's assumptions are violated and many of its significant results should not actually be considered significant, should we consider some alternative such as non-parametric or empirical tests? Not necessarily: non-parametric tests are making modeling assumptions or data transformations that are wrong as well; as statistician George E. P. Box famously said, "all models are wrong, but some are useful" [Box 1979]. *Every* model we will ever use to analyze system results is wrong, and all we can do is be aware of their shortcomings and try to improve them over time. The linear model is a very well-studied, very well-understood model; that provides many advantages.

We can always use empirical models. There is an "exact" permutation procedure for post hoc testing of multiple pairwise hypotheses [Pesarin 2001]. Like most permutation tests, it is implemented using random sampling. Algorithm 1 gives pseudo-code for a simple version. Specifically, this is a randomized version of Tukey's HSD that we described in Section 3.3.1. The algorithm forms a null distribution of the difference between the maximum MAP and the minimum MAP under the hypothesis that there is no difference between any of the systems. Each pairwise difference can then be compared against this distribution to determine whether it is "honestly" significant. This is a less powerful approach than the linear model plus single-step adjustment method we used above, but it requires fewer assumptions.

---

**Algorithm 1** Randomization test for all-pairs two-sided differences in MAP with  $p$ -values adjusted for multiple comparisons by Tukey's HSD.

---

**Require:**  $n \times m$  matrix  $\mathbf{X}$ , with each column vector  $\mathbf{X}_{\cdot i}$  containing APs for system  $i$  over  $n$  topics; number of trials  $B$

**Ensure:**  $p$ , a matrix of  $p$ -values for  $m(m-1)/2$  pairwise system comparisons

```

1: for each trial  $k$  in 1 to  $B$  do
2:   initialize  $n \times m$  matrix  $\mathbf{X}^*$ 
3:   for each topic  $t$  do
4:      $\mathbf{X}_{t \cdot}^* \leftarrow$  permutation of values in row  $t$  of  $\mathbf{X}$ 
5:   end for
6:    $q^* \leftarrow \max_i \bar{\mathbf{X}}_{\cdot i}^* - \min_j \bar{\mathbf{X}}_{\cdot j}^*$ 
7:   for each pair of systems  $i, j$  do
8:      $p_{ij} \leftarrow p_{ij} + 1/B$  if  $q^* > |\bar{\mathbf{X}}_{\cdot i} - \bar{\mathbf{X}}_{\cdot j}|$ 
9:   end for
10: end for
11: return  $p$ 

```

---

There are also non-parametric rank-based alternatives, just as the Wilcoxon signed rank test is a non-parametric rank-based alternative to the t-test. We will not cover them here, since we feel violations of the parametric assumptions usually



do not pose serious problems, and the randomization test is an excellent alternative if there is evidence that they do.

## 6. CONCLUSIONS

We have presented two issues in statistical analysis that affect our conclusions about significance between pairs of runs: the first is the data used to fit a model in which inferences about runs will be made, and the second is the Multiple Comparisons Problem and the adjustment of  $p$ -values to account for it. Using runs submitted to TREC, we showed that these two issues combined can have a radical effect on the way we interpret experimental data in the Cranfield paradigm: depending on how we define a family of experiments for MCP adjustment, it is possible that most of the pairs of systems considered significantly different by a paired t-test would no longer be significant once our adjustments are applied. At a high level our results support the conclusions of Armstrong et al. [2009b] that there is little reason to believe that there has been much improvement in ad hoc retrieval effectiveness.

There are still many questions without objective answers: How should significance be considered in publication decisions? How should a family of experiments be defined for  $p$ -value adjustment? How can learning effects on a given test collection be accounted for? All of these questions factor into the construction of a model within which to perform inferences about systems, and therefore into our understanding about IR system effectiveness. Furthermore, we have not even considered all the other ways multiple comparisons can happen: by looking at multiple evaluation measures, by using multiple hypothesis tests, by evaluating over multiple collections, and so on. All of these could be incorporated into a model for inference, but of course the more factors we model, the less likely we are to find significant differences—or the more annotated data we will need in order to find them. In short, MCP is a deep problem that colors all of the evaluation we do, particularly with reusable test collections.

With all of these questions, it is hard to escape the feeling that statistical inference is ultimately highly subjective, only providing a thin veneer of objectivity that makes us feel a little more comfortable about experimental rigor. That does not mean we should throw out the entire edifice—on the contrary, though we believe our analysis conclusively shows that a  $p$ -value cannot have any objective meaning, we still believe  $p$ -values from paired t-tests provide a decent rough indicator that is useful for many of the purposes they are currently used for. We only argue that  $p$ -values and significance test results in general should be taken with a very large grain of salt, and in particular have an extremely limited effect on publication decisions and community-wide decisions about “interesting” research directions.

In the long term, we believe that IR experimental analysis should transition to a fully-Bayesian modeling approach [Gelman et al. 2004]. In the Bayesian framework, we specify a model first, then fit it to data and perform inference. The advantage of this is that we make all modeling assumptions explicit; even if the initial models are incorrect, over time they can be refined in a community effort to handle the specific challenges of information retrieval evaluation. Gelman et al. [2009] have shown that a Bayesian approach can have a strong mitigating effect on MCP; this is one example of the advantage it provides.

## REFERENCES

- ARMSTRONG, T. G., MOFFAT, A., WEBBER, W., AND ZOBEL, J. 2009a. Evaluation: Measurement and certification of IR systems. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 25–26.
- ARMSTRONG, T. G., MOFFAT, A., WEBBER, W., AND ZOBEL, J. 2009b. Improvements that don't add up: ad hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*.
- BERGER, J. O. 2003. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science* 18, 1, 1–32.
- BOX, G. E. P. 1979. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, R. Launer and G. Wilkinson, Eds.
- BRETZ, F., HOTHORN, T., AND WESTFALL, P. 2010. *Multiple Comparisons Using R*, 1st ed. CRC Press.
- BUCKLEY, C. AND VOORHEES, E. M. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 33–40.
- CORMACK, G. V. AND LYNAM, T. R. 2006. Statistical precision of information retrieval evaluation. In *Proceedings of SIGIR*. 533–540.
- CORMACK, G. V. AND LYNAM, T. R. 2007. Power and bias of subset pooling strategies. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, New York, NY, USA. To appear.
- DRAPER, N. R. AND SMITH, H. 1998. *Applied Regression Analysis*, 3rd ed. Wiley-Interscience.
- GELMAN, A., CARLIN, J. B., STERN, H. S., AND RUBIN, D. B. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC.
- GELMAN, A., HILL, J., AND YAJIMA, M. 2009. Why we (usually) don't have to worry about multiple comparisons. arXiv:0907.2578v1 [stat.AP].
- HOTHORN, T., BRETZ, F., AND WESTFALL, P. 2008. Simultaneous inference in general parametric models. *Biometrical Journal* 50, 3, 346–363.
- HSU, J. 1996. *Multiple Comparisons: Theory and Methods*, 1st ed. Chapman and Hall/CRC.
- HULL, D. A. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR*. 329–338.
- IOANNIDIS, J. P. A. 2005a. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* 294, 2, 218–228.
- IOANNIDIS, J. P. A. 2005b. Why most published research findings are false. *PLoS Medicine* 2, 8.
- MARDIA, K. V., KENT, J. T., AND BIBBY, J. M. 1980. *Multivariate Analysis*, 1st ed. Academic Press.
- MCCULLAGH, P. AND NELDER, J. 1989. *Generalized Linear Models*, 2nd ed. Chapman and Hall/CRC.
- MILLER, R. G. 1981. *Simultaneous Statistical Inference*, 2nd ed. Springer.
- MONAHAN, J. F. 2008. *A Primer on Linear Models*, 1st ed. Chapman and Hall/CRC.
- PESARIN, F. 2001. *Multivariate Permutation Tests: With Applications in Biostatistics*, 1st ed. Wiley.
- RAUDENBUSH, S. W. AND BRYK, A. S. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Sage.
- SANDERSON, M. AND ZOBEL, J. 2005. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of SIGIR*. 186–193.
- SAVOY, J. 1997. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management* 33, 4, 495–512.
- SMUCKER, M., ALLAN, J., AND CARTERETTE, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM*. 623–632.
- SMUCKER, M. D., ALLAN, J., AND CARTERETTE, B. 2009. A comparison of statistical significance tests and small sample sizes for IR evaluation. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 630–631.

- VAN RIJSBERGEN, C. J. 1979. *Information Retrieval*. Butterworths, London, UK.
- VENABLES, W. N. AND RIPLEY, B. D. 2002. *Modern Applied Statistics with S*, 4th ed. Springer.
- VOORHEES, E. M. AND HARMAN, D. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. 1–24. NIST Special Publication 500-246.
- VOORHEES, E. M. AND HARMAN, D. K., Eds. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.
- WALD, A. 1947. *Sequential Analysis*. John Wiley, New York.
- WASSERMAN, L. 2003. *All of Statistics*. Springer.
- WASSERMAN, L. 2006. *All of Nonparametric Statistics*. Springer.
- WILBUR, W. J. 1994. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Sciences* 20, 4, 270–284.
- ZOBEL, J. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments? In *Proceedings of SIGIR*. 307–314.

## A. LINEAR MODELS IN R

It is instructive to check the equivalencies we claim in Section 2.1 using a statistical toolkit such as R. We will assume that we have an R data frame in which each row has a measurement  $y_{ij}$ , a categorical system identifier, and a categorical topic number. For example:

```
R> X
      y system topic
1  0.0273   S1  401
2  0.1131   S1  402
3  0.7415   S1  403
...
50 0.3094   S1  450
51 0.0067   S2  401
52 0.2094   S2  402
53 0.9124   S2  403
...
100 0.1208   S2  450
```

To perform a one-sided paired t-test between  $S_1$  and  $S_2$ , we use the `t.test` procedure:

```
R> t.test(y ~ system, paired=T, data=X)
...
t = -1.1653, df = 49, p-value = 0.2495
...
```

We could also use a within-subjects ANOVA with block errors on topics:

```
R> summary(aov(y ~ system + Error(topic), data=X))
...
Error: Within
      Df Sum Sq Mean Sq F value Pr(>F)
system  1 0.006927 0.0069272  1.3579 0.2495
Residuals 49 0.249962 0.0051013
```

Note that the  $p$ -values returned by the `t.test` and the ANOVA are identical. The square of the  $t$  statistic is equal to the  $F$  statistic reported by `aov`.

The `lm` procedure fits a linear regression model:

```
R> summary(lm(y ~ system + topic, data=X))
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.008677   0.051006   0.170  0.865619
systemS2     0.016646   0.014285   1.165  0.249535
...
```

We can see that the  $t$  statistic on the coefficient corresponding to  $S_2$  reported by `lm` is the absolute value of that reported by `t.test`, and the  $p$ -values are identical. We also note that the standard error on that coefficient is a function of the residual mean square error reported by `aov`:  $0.014285 = \sqrt{2/50 \cdot 0.0051013}$ . These equivalencies will hold no matter what data we use; it need not conform to any of the assumptions of the tests.

Finally, we use `lmer` in the `lme4` package. `lmer` is capable of much greater modeling power than any of `t.test`, `aov`, or `lm`, though the details are beyond the scope of this work.

```
R> lmer(y ~ system + (1|topic), data=X)
...
Random effects:
 Groups   Name      Variance Std.Dev.
topic    (Intercept) 0.0359778 0.189678
Residual                    0.0051013 0.071423
...
Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.23254    0.02866   8.113
systemS2     0.01665    0.01428   1.165
...
```

Though the intercept estimate is different from `lm`, the quantities we are most interested in—the standard error  $s_\beta$  and  $t$  statistic for the coefficient on  $S_2$  and the residual variance  $\hat{\sigma}^2$ —are identical to those reported by the other procedures.

## B. MULTIPLE TESTING TUTORIAL FOR R

This section provides a tutorial on performing the analysis in Sections 2 and 3 using the statistical programming environment R.

### B.1 Example Experimental Evaluation

In this scenario, a researcher is comparing five systems: a baseline  $S_0$ , two approaches that use the same model but different feature sets  $S_1, S_2$ , and two approaches that use a more advanced model with the same respective feature sets  $S'_1, S'_2$ . All systems are run over the same topics. The researchers want to know

which feature set is better and which model is better, and compare everything to the baseline; thus the suite of one-sided hypotheses to be tested is:<sup>6</sup>

$$\begin{array}{ll} S_1 \leq S_0 & S'_1 \leq S_1 \\ S_2 \leq S_0 & S'_2 \leq S_2 \\ S'_1 \leq S_0 & S_1 \leq S_2 \\ S'_2 \leq S_0 & S'_1 \leq S'_2 \end{array}$$

We want to avoid testing all 10 pairwise hypotheses, because the more we test, the less inferential power we have for each individual test. It is prudent to look at the data and decide in advance which hypotheses to test. Suppose we run the evaluation and note that  $S'_1$  has a greater mean than  $S_1$  and  $S'_2$  has a greater mean than  $S_2$ . In that case we can drop the hypotheses  $S'_1 > S_0$  and  $S'_2 > S_0$ , since in our modeling regime test results are transitive. Therefore we have 6 hypotheses to test about 5 systems.

We fit a linear model with system effects and topic effects. Assume we have the data in an R data frame, with one row for each system/topic pair:

```
R> res
      y system topic
1 0.0897     S0     1
2 0.0000     S0     2
...
51 0.2830     S1     1
52 0.0168     S1     2
...
```

We first fit a linear model to the data frame using the `lmer` function in the `lme4` package. As we showed in Section A, `lm` would work as well; we chose to use `lmer` because it provides access to a wider and more powerful range of analytical tools.

```
R> lmod <- lmer(y ~ system+(1|topic), data=res)
```

Once we have the model, we need to test the omnibus hypothesis that  $S_0 = S_1 = S_2 = S'_1 = S'_2$ . If we cannot reject that, there is no point in continuing with the analysis.

```
R> anova(lmod)
Analysis of Variance Table
      Df Sum Sq Mean Sq F value
system 4 0.67964 0.16991  20.255
```

An  $F$  statistic of 20 is very high and unlikely to occur by chance. We can confirm that it is significant by looking it up in a table of  $F$  values. Unfortunately the `lmer` routines do not provide residual degrees of freedom, but since our design is fully nested we can calculate them as  $(n - 1)(m - 1)$ , or  $49 \cdot 4 = 196$  for this example. Then we obtain a  $p$ -value:

<sup>6</sup>These could also be two-sided hypotheses with no change to most of this section.

```
R> pf(20.255, df1=4, df2=196, lower.tail=F)
[1] 5.58735e-14
```

We can conclude that at least two of the systems are significantly different. We do not yet know which.

Looking at the model in more detail, we see the following output:

```
R> lmod
Linear mixed model fit by REML
Formula: y ~ system + (1 | topic)
Data: res
      AIC      BIC logLik deviance REMLdev
-409.8 -385.1 211.9  -423.8  -390.8
Random effects:
Groups   Name             Variance Std.Dev.
topic    (Intercept) 0.0041181 0.064172
Residual                    0.0083886 0.091589
Number of obs: 250, groups: topic, 50

Fixed effects:
              Estimate Std. Error t value
(Intercept)   0.02865    0.01582   1.811
systemS1      0.03380    0.01832   1.845
systemS1'     0.05365    0.01832   6.341
systemS2      0.11616    0.01832   2.929
systemS2'     0.14064    0.01832   7.678
```

The  $t$  statistics can be used to test hypotheses about the difference between each system and the baseline. Note that  $p$ -values are not provided—this is partially because those  $p$ -values would not be adjusted for multiple comparisons. We can also test for differences between systems; the  $t$  statistic for the difference between  $S_1$  and  $S_1'$  would be  $t = (0.05365 - 0.03380)/0.01832$ . Again, however,  $p$ -values for those will not be adjusted.

To set up the six hypotheses of interest for  $p$ -value adjustment, we need to create a contrast matrix  $K$ . The number of columns of  $K$  is  $m$ , the number of systems (in this case  $m = 5$ ), and the number of rows is the number of hypotheses to be tested. Each row represents a hypothesis and each column represents a system.

```
R> K
      (Intercept) S1 S1' S2 S2'
S1 - S0           0  1  0  0  0
S2 - S0           0  0  0  1  0
S1' - S1          0 -1  1  0  0
S2' - S2          0  0  0 -1  1
S2 - S1           0 -1  0  1  0
S2' - S1'         0  0 -1  0  1
```

The first two rows request a test between  $S_1$  and  $S_0$  and between  $S_2$  and  $S_0$ .  $S_0$  is not actually represented as a column since it is subsumed by the model intercept.

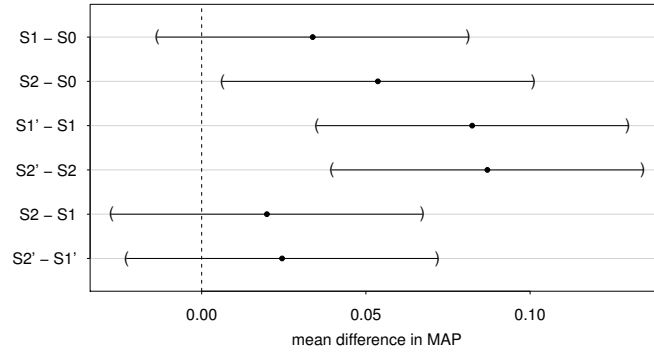


Fig. 11. 95% adjusted confidence intervals for the six hypotheses under consideration. The three intervals that do not contain 0 are the three hypotheses we reject.

The next four rows request tests between pairs of non-baseline systems. Each row has a 1 and a -1 to indicate the systems involved in the test. The signs indicate which direction to test in a one-sided test; in a two-sided test they do not matter.

To test all six hypotheses, we use the `glht` routine in the `multcomp` package. The default  $p$ -value adjustment procedure is the single-step multivariate  $t$  distribution approach we described in Section 3.3.2.

```
R> summary(glht(lmod, linfct=K, alternative="greater"))
```

#### Simultaneous Tests for General Linear Hypotheses

```
Fit: lmer(formula = y ~ system + (1 | topic), data = res)
```

Linear Hypothesis:

		Estimate	Std. Error	z value	Pr(>z)	
S1	- S0	<= 0	0.03380	0.01832	1.845	0.16317
S2	- S0	<= 0	0.05365	0.01832	2.929	0.00974 **
S1'	- S1	<= 0	0.08236	0.01832	4.496	< 0.001 ***
S2'	- S2	<= 0	0.08699	0.01832	4.749	< 0.001 ***
S2	- S1	<= 0	0.01985	0.01832	1.084	0.54779
S2'	- S1'	<= 0	0.02448	0.01832	1.337	0.39563

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

(Adjusted p values reported -- single-step method)

This tells us that  $S_2$  is significantly better than the baseline, that  $S_1'$  is significantly better than  $S_1$ , and that  $S_2'$  is significantly better than  $S_2$ , but none of the other tests are significantly different. This suggests that the retrieval model matters but the features do not. Note that we can infer that  $S_1'$  is significantly better than the baseline even though we did not test for it; because the coefficient on  $S_1'$  is greater than the coefficient on  $S_2$  and we rejected  $S_2 \leq S_0$ , we can also reject  $S_1' \leq S_0$  by transitivity. The use of a model fit to all data allows this kind of inference which

is invalid in the independent all-pairs regime.

We can also look at adjusted confidence intervals for the differences in means for each of our hypotheses. The `confint` function prints them; plotting a `glht` object shows them in a graph. Figure 11 shows the 95% adjusted confidence intervals for our six hypotheses.

Finally, we note that had we done the typical unadjusted independent all-pairs t-tests, we would have found  $S_1$  significantly better than  $S_0$  and  $S_2$  significantly better than  $S_1$ . Taking into account interdependence and multiple comparisons reveals that we do not have enough evidence to conclude either of those with high confidence.

## B.2 Writing results for publication

We have detailed all steps involved in testing multiple hypotheses about pairs of systems. While it seems complicated, it can be performed as if it were a “black box” and written up straightforwardly. A paper submitted for publication would only need a paragraph such as this:

We performed an ANOVA to test the omnibus null hypothesis that  $S_0 = S_1 = S_2 = S'_1 = S'_2$  by mean average precision (MAP). This was rejected with  $p < 0.05$  ( $F = 20$  on 4 system d.f. and 196 residual d.f.), meaning at least two of the systems are significantly different. We then performed post hoc pairwise t-tests for  $S_1 \leq S_0$ ,  $S_2 \leq S_0$ ,  $S'_1 \leq S_1$ ,  $S'_2 \leq S_2$ ,  $S_1 \leq S_2$ , and  $S'_1 \leq S'_2$ , using the single-step method to adjust the  $p$ -values for multiple comparisons. The results are shown in Table 1.

And a table like this:

system	model	feature set	MAP
$S_0$	baseline	LM	0.1786
$S_1$	M1	PRF	0.2124
$S_2$	M2	PRF	0.2323*
$S'_1$	M1	MRF	0.2948*†
$S'_2$	M2	MRF	0.3193*†

Table 1: (Example table of results for a publication.) Results of our experiments on 50 TREC topics. \* indicates significance over the baseline. † indicates significance of MRF features over PRF features. No significant differences were found between M1 and M2. All tests are post hoc one-sided t-tests at the  $p < 0.05$  level.

We do not recommend reporting the adjusted  $p$ -values themselves (since there are so many sources of variance in estimating them), but we recognize that there are times when it can be helpful to do so.

## B.3 Code

We have written some R functions to perform analysis for common scenarios in IR experimentation. A function `pairwise.test` takes evaluation data from multiple systems, fits a model to the specified measure (which must be one of the ones output by `trec.eval`), and reports pairwise test results for specified hypotheses. For our example above, the researcher could do the following:



```
R> res <- read.eval("evals/", names=c("S0", "S1", "S2", "S1'", "S2'"))
R> pairwise.test(res, "map",
  H=c("S1 - S0", "S2 - S0", "S1' - S1",
      "S2' - S2", "S1 - S2", "S1' - S2'"),
  alternative="greater")
```

The `read.eval` function reads output from `trec_eval -q` and gives each system a name. The `pairwise.test` function fits a model, creates the contrast matrix for the requested pairwise tests, and then performs the tests. Paired two-sided tests are done by default.