Uncontextualized significance considered dangerous

Nicola Ferro
University of Padua
Padova, Italy
nicola.ferro@unipd.it

Mark Sanderson
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

ABSTRACT

We examine the context of significance tests in offline retrieval experiments. Our *Information Retrieval (IR)* community is notable for its experimental rigour: the use of statistical significance is grows across our publications. However, we show that ignoring the context of a test risks Type I errors, leading to potential publication bias. We examine two contexts: multiple testing and the types of the retrieval systems being compared. Our results show that multiple testing corrections are critical for experimental work. In addition, we find that past research on the reliability of test collections maybe flawed owing to the type of systems examined. The latter result has not been shown before. Together our results suggest substantial numbers of Type I errors in offline IR experiments. We detail a methodology to alleviate the errors.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Retrieval effectiveness**.

KEYWORDS

statistical significance testing; comparing tests; ANOVA; prediction

ACM Reference Format:
Nicola Ferro and Mark Sanderson. 2024. Uncontextualized significance considered dangerous. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657827

1 INTRODUCTION

Test collections are a mainstay of offline evaluation in IR [32]. Their measurement accuracy became a focus of attention starting with work by Zobel [43], which led to a series of papers that cumulatively amassed evidence indicating test collection measurements were reliable. Zobel created a topic splitting methodology to assess result consistency, Zobel also examined the properties of a number of statistical significance tests showing their value in measurement. Topic splitting and significance became two key statistical pillars used to substantiate the experiments of IR research.

Combining topic splitting with the popular t-test, Ferro and Sanderson [13] reported "*an unexpected number of inconsistent [measurements]*": statistically significant results were occurring due to chance (i.e. *Type I errors*) far more than past work [33, 43] anticipated. Ferro and Sanderson stated the inconsistency could lead to *publication bias*: where a paper is more likely to be accepted if it reports a statistically significant improvement [9, 11]. The inconsistency is important to understand, as Ferro and Sanderson claimed that their topic splitting experiments were more representative of actual IR experiments than those pioneered by Zobel. In this paper, therefore, we address the following research questions:

- Is there a prevalence of Type I errors in IR experiments of the type typically published in IR forums?
- How does the error vary across different types of IR experiments?
- If such variation occurs, how best can it be alleviated?

We find that there is a need to understand the *context* of a significance test: was the test part of a family of multiple comparisons, and what types of IR systems are being compared?

2 PAST WORK

We detail past work in topic splitting and the use of significance.

2.1 Topic splitting

Zobel [43] created a topic splitting methodology to assess test collection result consistency. The methodology split the topics of the TREC 5 ad hoc track collection into two equal sized sets. Zobel pairwise compared, over both sets, the output of every system submitted to the track, i.e. the *runs*. Across the "*approximately 4500*" pairs, Zobel noted, in one topic set, comparisons that resulted in a significant improvement of one run u over another v ($u \succ\succ v$). He then examined if $u \leq v$ in the other set, reporting that across the 4,500 comparisons there were only 7 inconsistent pairs. Zobel concluded that test collection experiments "*lead to reliable results*".

Voorhees and Buckley [40] extended Zobel's methodology by randomly splitting topic sets fifty times across multiple test collections and discarding the 25% least effective runs "*to prevent these uninteresting runs from having an effect on our calculations*", an approach widely adopted by others [12, 13, 33, 38].[1] Voorhees and Buckley also restricted pairwise comparisons to a particular type of run, those submitted from a single participant. The authors stated that "*While runs submitted to TREC by the same [participant] are not necessarily variants of a common system, very frequently they are*". They reported the participant comparisons were more consistent (having "*lower error rates*") compared to runs across the whole track. The paper is one of the few times that measurement accuracy participant runs was tested.

SIGIR '24, July 14–18, 2024, Washington, DC, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657827

[1] See also Boytsov et al. [6] for a notable variation on the methodology using a collection with tens of thousands of topics.

2.2 Significance

The steady adoption of significance in IR experiments is detailed in past reviews [29, 32]. While many researchers use significance, often the tests are conducted multiple times without correction. This practice has been criticized as greatly increasing the probability of Type I errors [16, 31]. There have been empirical examinations of this mistake. Blanco and Zaragoza [4] examined simulated results and Boytsov et al. [6] examined which might be the best correction method. However, the relationship between correction methods and the nature of IR experiments, such as topic set size or the types of runs used has not been examined.

3 METHODOLOGY

We study how statistical significance tests behave when comparing different types of runs. We consider the following set conditions:

- **track**: all the runs submitted to the track of an evaluation campaign are considered. Typically 50–150 runs with approximately 1,200-11,000 pair-wise comparisons. This is the context organizers of a track use when summarizing the results of submissions and is typically adopted by IR researchers when developing a new evaluation measure, pooling strategy, studying a statistical significance test, etc.
- **participant**: just the runs submitted by a participant to a track of an evaluation campaign are considered. Around ten runs amounting to (roughly) 50 pair-wise comparisons. This is the context a typical IR researcher might find themselves trying to decide which version of their system is better. As detailed in Sec 2 this context is overlooked. Analyses conducted in the track condition are assumed to hold also for the participant one.

In both conditions, comparing all possible pairs of runs risks inflating *Type I error* probability, i.e. the probability of rejecting the null hypothesis and considering two runs as significantly different when they are not. If α represents the Type I error probability when comparing one pair of runs, when performing m independent pair-wise comparisons, the probability of committing at least one Type I error in the m comparisons becomes $1 - (1 - \alpha)^m$; the so-called *Family-wise Error Rate (FWER)* [17, 18]. *Multiple comparisons* require adjustment of a significance test so as to avoid inflating the Type I error probability.

3.1 Statistical Significance Tests

The methodology we describe can be applied to any statistical significance test, here, we focus on parametric tests: the Student's t test [35] and *ANalysis Of VAriance (ANOVA)* [15]. The former allows comparison of two runs at a time and it is still the most used statistical significance test by IR practitioners [29]; the latter allows comparing a set of runs together. It can be considered as a generalization of the Student's t test to multiple runs.

Non parametric significance tests, such as the Wilcoxon test [42], are also adopted by IR practitioners and some authors Parapar et al. [23, 24]. The methodology proposed in this paper can be applied to non parametric significance tests. We leave for future work a derived formal analysis, which would require use of ranks instead of marginal means as the one for parametric tests in this Section, and to conduct experiments.

Let us consider a set of $i = 1, \dots, T$ topics and $j = 1, \dots, R$ runs. Let y_{ij} be the performance, according to some evaluation measure, of run j on topic i. Since we aim at comparing all the possible pairs of runs, we need to perform $m = \binom{R}{2} = \frac{R(R-1)}{2}$ comparisons.

3.1.1 *Student's t test.*

The test compares two runs u and v. We use a paired t test, since each topic is applied to both runs.

Let $d_i = y_{iu} - y_{iv}$ the difference between the performance of the two runs u and v on topic i and let $\hat{\mu}_d$ and $\hat{\sigma}_d^2$ be the sample mean and the sample variance of the performance difference over the T topics. The null hypothesis H_0 is that the performance of the two runs u and v is the same, i.e. the population mean is $\mu_d = 0$. Under the null hypothesis, the test statistic

$$t_{stat} = \frac{|\hat{\mu}_d|}{\sqrt{\hat{\sigma}_d^2 / T}} > t_{T-1}^{1-\alpha/2} \tag{1}$$

is distributed as a Student's t distribution with $T - 1$ degrees of freedom and, for a two-tailed test – i.e. run u is greater than run v or vice-versa – its value has to be above the $100 * (1 - \alpha/2)$-th percentile of the Student's t distribution in order to reject the null hypothesis; note that we use $\alpha/2$ since the Student's t distribution is symmetric with respect to the origin and we are conducting a two-tailed test. The Student's t distribution allows us to compute the p-value $2 \cdot \mathbb{P}\left(t_{T-1} \geq t_{stat} | H_0\right)$ (for a two-tailed test) and verify that $p \leq \alpha$ to decide whether to reject the null hypothesis or not.

The test in eq. (1) focuses on two runs in isolation. Consequently, its value t_{stat} – or, equivalently, the p-value – will be the same in both *track* and *participant* conditions.

In order to compare all possible pairs of runs, we need to perform m separate t tests. To control for the FWER, we adopt Bonferroni's correction [5] which adjusts the significance level as follows:

$$\alpha' = \frac{\alpha}{m} \tag{2}$$

The adjusted significance level α' can be used to compute a different threshold $t_{T-1}^{1-\alpha'/2}$ in eq. (1) or, equivalently, to be compared against the p-value of each t test.

The Bonferroni's correction (2) does not consider information about the two runs being compared or about the whole set of runs R, it only relies on the fact that m comparisons have to be performed. However, since the total number of comparisons m is different in the *track* and in the *participant* conditions, the adjusted α' significance level will be different in the two conditions. In other terms, while the test statistics in eq. (1) produce the same t_{stat} value in both the *track* and *participant* conditions and, consequently, the same p-value, this t_{stat} value will be compared against two different $t_{T-1}^{1-\alpha'/2}$ thresholds (α') in the *track* and *participant* conditions.

We label the paired t-test with Bonferroni's correction `ttpB`. Since it is erroneous but common in IR to conduct multiple t tests without FWER adjustment, we also consider uncorrected Non-Bonferroni tests, labelled `ttpNB`. In the `ttpNB` case, t_{stat} is compared against the same (un-adjusted) threshold $t_{T-1}^{1-\alpha/2}$. In other terms, the p-value from the test statistics is compared against the same (un-adjusted) α in both the *track* and the *participant* conditions. Therefore, not performing the Bonferroni's correction not only inflates the Type I error rate but it also wrongly removes any difference between the *track* and *participant* conditions.

3.1.2 Two-way ANOVA. We use the following model [21, 28]

$$y_{ij} = \mu_{\cdot\cdot} + \beta_i + \gamma_j + \varepsilon_{ij} \tag{3}$$

where $\mu_{\cdot\cdot}$ is the grand mean, β_i is the effect of the i-th topic, γ_j is the effect of the j-th run, and ε_{ij} is the residual error committed by the model in predicting y_{ij}. Two-way ANOVA allows consideration of all R runs together; note that, when used for just two runs, it is substantially equivalent to the paired Student's t test. This model was used for the first time by Tague-Sutcliffe and Blustein [36] to analyze TREC 3 data, then by Banks et al. [2] to conduct a more extensive analysis on TREC data and, eventually, studied and extended by others [12–14, 27, 41].

The two-way ANOVA tests the so-called omnibus null hypothesis that all the R runs perform the same; rejecting the null hypothesis means that at least one run should perform differently from others. As we are interested in all the m pair-wise comparisons across R, we need to pair the two-way ANOVA with a follow-up Tukey *Honestly Significant Difference (HSD)* test [37], which controls for the FWER. The Tukey HSD test uses the following test statistic:

$$t_{stat} = \frac{|\hat{\mu}_{\cdot u} - \hat{\mu}_{\cdot v}|}{\sqrt{\mathrm{MS_{Error}}/T}} > q^{\alpha}_{R,\mathrm{df_{Error}}} \tag{4}$$

where: $\hat{\mu}_{\cdot u} = \hat{\mu}_{\cdot\cdot} + \hat{\gamma}_u$ and $\hat{\mu}_{\cdot v} = \hat{\mu}_{\cdot\cdot} + \hat{\gamma}_v$ are the marginal means of u and v as estimated from the actual data; $\mathrm{df_{Error}}$ are the *Degrees of Freedom (DF)* of the error $\hat{\varepsilon}_{ij}$; $\mathrm{MS_{Error}}$ is the *Mean Squares (MS)* of the error, i.e., an estimation of the variance left unexplained by the ANOVA model; and $q^{\alpha}_{R,\mathrm{df_{Error}}}$ is the upper $100 * (1 - \alpha)$-th percentile of the studentized range distribution [22], from which we compute the p-value $\mathbb{P}\left(q_{R,\mathrm{df_{Error}}} \geq t_{stat} | H_0\right)$ using $p \leq \alpha$ to decide null hypothesis rejection/acceptance. Note the p-value already accounts for the adjustment for the multiple comparisons and can be compared to the significance level α without further correction.

Note that eq. (1) and eq. (4) both normalize the mean performance difference between two runs by a quantity proportional to a standard deviation and compare the difference against an appropriate threshold. Eq. (1) focuses on the two runs examined and normalizes their mean performance difference by a quantity proportional to their standard deviation, a quantity which changes for each pair of runs. Eq. (4) contextualizes the difference with respect to the whole set R of considered runs. This is done because the mean performance difference is normalized by the "standard deviation" of the error of the two-way ANOVA model fitted on R, which is the same for each pair of runs. When running multiple t tests, the threshold $t^{1-\alpha/2}_{T-1}$ is the same for each pair of runs but independent from all of them, since it is a function of only on the number of topics T used and the significance level α. On the other hand, the studentized range distribution (the threshold $q^{\alpha}_{R,\mathrm{df_{Error}}}$) is parameterized by both the degrees of freedom of the error in the ANOVA model and the total number of runs R under comparison (besides the significance level α). Therefore, it is same for all the pairs of runs under examination but it takes into account the whole set of runs rather than being independent from them.

While the mean performance difference between the runs u and v is the same in both the *track* and the *participant* conditions, both the normalization factor and the studentized range distribution are different in the *track* and in the *participant* conditions, thus adapting both the test statistics and the threshold to each condition.

It is incorrect to conduct a Tukey HSD test only when ANOVA rejects the omnibus null hypothesis (see Hsu [18, p. 177ff] and Sakai [30, p. 73] for details). We always perform the Tukey HSD test, regardless of the outcome of the omnibus null hypothesis.

We label two-way ANOVA with Tukey HSD as `anv2`.

3.2 Run Set-Wise Counts

3.2.1 Topic Splitting. To study statistical significance tests across a set of runs, we adopt the *topic splitting* methodology used by Ferro and Sanderson [13]. We sample without replacement from T to form two equal size topic sets – TS1 and TS2 – of varying sizes. We compute the statistical significance tests under – `ttpB`, `ttpNB`, and `anv2` – on each TS and we consider the level of agreement on the two topic sets TS1 and TS2 according to several measures. We use the same TS splits for all the significance tests and conditions (*track* or *participant*). This allows us to understand if the conclusions we draw from one experiment, hold for another.

For each TS size, we repeat this whole process $S = 1,000$ times by re-sampling the TS aggregating performance across the samples.

Note that eq. (1) can be re-written as $t_{stat} = \frac{\hat{\mu}_d}{\sqrt{\hat{\sigma}^2_d}} > \frac{1}{\sqrt{T}} t^{1-\alpha/2}_{T-1}$ and eq. (4) as $t_{stat} = \frac{|\hat{\mu}_{\cdot u} - \hat{\mu}_{\cdot v}|}{\sqrt{\mathrm{MS_{Error}}}} > \frac{1}{\sqrt{T}} q^{\alpha}_{R,\mathrm{df_{Error}}}$. The threshold against which the normalized performance difference between runs is compared against is the same not only for each pair of runs under comparison (as discussed in the previous section) but also for any two topic sets TS1 and TS2 of the same size and for all the S samples of topic sets. Therefore, the topic splitting methodology allows us to set a reference (the threshold), which is common to all the experiments, ensuring a fair level of comparability.

3.2.2 Consistency Counts. From Ferro and Sanderson [13], we consider the following axes:

- *Significance*: what a statistical significance test detects on the two TSs, broken down into: **Active (A)**, both TSs say "significantly different"; **Passive (P)**, both TSs say "not significantly different"; **Mixed (M)**, one TS says "significantly different" but the other set says "not significantly different".
- *Order*: how runs are ordered on the two TSs; either **Agreement (A)**, both TSs say that one run is better than another ($u \succ v$), or viceversa; **Disagreement (D)**, one TSs says $u \succ v$ but the other set says $u \prec v$, or viceversa.

From this, we can sub-divide what happens to each pair of runs (u, v) on two TSs as follows:

- **Active Agreement (AA)**: on both TS1 and TS2, either $u \succ\succ v$ or $v \succ\succ u$. The best kind of consistency.
- **Active Disagreements (AD)**: $u \succ\succ v$ on TS1 but $v \succ\succ u$ on TS2; or $v \succ\succ u$ on TS1 but $u \succ\succ v$ on TS2. The worst kind of inconsistency, we reach opposite conclusions on the two TSs.
- **Mixed Agreement (MA)**: $u \succ\succ v$ on TS1 but $u \succ v$ on TS2; or $v \succ\succ u$ on TS1 but $v \succ u$ on TS2; $u \succ v$ on TS1 but $v \succ\succ v$ on TS2; or $v \succ u$ on TS1 but $v \succ\succ u$ on TS2. A situation where a test is not able to confirm its conclusions on both TSs. With the order of the two runs the same, however, it could indicate a lack of power in the significance test.
- **Mixed Disagreement (MD)**: $u \succ\succ v$ on TS1 but $v \succ u$ on TS2; or $v \succ\succ u$ on TS1 but $u \succ v$ on TS2; or $u \succ v$ on TS1 but $v \succ\succ u$

on TS2; or $v \succ u$ on TS1 but $u \succ\succ v$ on TS2. This indicates a situation where a test is not able to confirm its conclusions on both TSs and the order of the two runs is the opposite. More severe issue than MA but less than AD.

- **Passive Agreement (PA)**: on both TS1 and TS2, $u \succ v$ or $v \succ u$. A less important count given the lack of significance in either TS, but if it is too big, it could be also a symptom of lack of power.
- **Passive Disagreement (PD)**: $u \succ v$ on TS1 but $v \succ u$ on TS2; or $v \succ u$ on TS1 but $u \succ v$ on TS2. Disagreement that is a less severe than MD given nothing is significantly different.

The six counts summed over all the possible pairs of runs is equal to the total number of pairs under comparison

$$\sum_{(u,v)} (AA + AD + MA + MD + PA + PD) = m = \frac{R(R-1)}{2} \quad (5)$$

thus providing a complete account of what happens when comparing a whole set of runs. Note that similar counts in previous work on topic splitting [33, 38, 40, 43] do not hold this property and they do not sum up to the total number of run pairs under comparison.

3.2.3 Bias and Disagreement Rate. Ferro and Sanderson [13] also measure *Bias*: the likelihood of obtaining a significant result when a test on a different TS would have produced either no significance (MA, MD) or a significant result in the opposite direction (AD):

$$Bias = 1 - \frac{AA}{AA + AD + \frac{MA}{2} + \frac{MD}{2}} \quad (6)$$

The count of MA and MD is halved as in only one of the two TSs in those counts significance was observed. One minus the fraction is computed to focus on errors, the lower, the better.

We also consider the *Disagreement Rate*:

$$DR = \frac{AD + MD + PD}{AA + AD + MA + MD + PA + PD} = \frac{AD + MD + PD}{m} \quad (7)$$

which quantifies how much runs are in the opposite order in the two TSs and, thus, the lower, the better. It corresponds to the error or swap rate detailed in past work [33, 40].

Note that while *Bias* summarizes and accounts for the joint viewpoint of a significance test (significance axis) and how actual data layout runs (order axis), the *Disagreement Rate* accounts only for how actual data layout runs (order axis) and it is independent from the statistical significance test at hand.

3.3 Run Pair-Wise Probabilities

While Ferro and Sanderson focused on analyzing sets of runs as a whole, we extend their methodology to provide more insights on each pair of runs (u, v) under comparison.

Each run pair (u, v) can result in one outcome: AA, AD, MA, MD, PA, PD. We can model the AA case as a *binomial random variable* $X_{AA} \sim B(1, p_{AA})$ with parameters 1 and p_{AA}, where p_{AA} is the probability that the pair (u, v) is an AA. We can proceed with similar binomial random variables $X_{AD}, X_{MA}, X_{MD}, X_{PA}$, and X_{PD} and their corresponding probabilities for the other cases. To estimate p_{AA} (and the other cases), we rely on $\mathbb{E}[X_{AA}] = p_{AA}$, i.e. the expectation of a binomial random variable $B(1, p_{AA})$ is p_{AA}. As explained above, the topic splitting process is repeated S times, we can estimate the sample mean and thus p_{AA} as follows:

$$\hat{p}_{AA} = \frac{\sum_{i=1}^{S} \chi_{AA}(i)}{S} \quad (8)$$

where χ_{AA} is the indicator function denoting whether the run pair (u, v) is an AA in the i-th trial of the topic splitting process.

Note that, since AA, AD, MA, MD, PA, PD are disjoint outcomes covering the whole sample space Ω, it holds:

$$p_{AA} + p_{AD} + p_{MA} + p_{MD} + p_{PA} + p_{PD} = 1 \quad (9)$$

Thus, we can provide each pair of runs (u, v) with a fine-grained analytics on its probability to turn out to be an AA, AD, etc. Thus we can provide an estimate of the "reliability" of the run pair (u, v), helping researchers and practitioners in taking decisions about that pair. For example, suppose u is a new version of our IR system and v is the version currently in operation, if for the pair (u, v) the p_{AA} was high enough and $u \succ\succ v$, one might decide to put u in production instead of v. However, if p_{AD} was not low enough (or just above zero), we could take the opposite decision and keep v in production, since the results are not stable enough yet.

We can also exploit these probabilities and the fact that they are disjoint events to estimate the probability that a run pair (u, v) would contribute to the bias

$$p_{Bias} = p_{AD} + p_{MA} + p_{MD} \quad (10)$$

or to the disagreement rate

$$p_{DR} = p_{AD} + p_{MD} + p_{PD} \quad (11)$$

3.4 Topic Splitting With/Without Replacement

Sanderson and Zobel [33] observed that, as the TS size approaches $\frac{T}{2}$, the splits become less independent since sampling is without replacement. Indeed, when $ts_1 = ts_2 = \frac{T}{2}$, once the topics in TS1 are sampled, the topics in TS2 are already chosen. Sanderson and Zobel concluded that this dependency might artificially inflate what we called *Disagreement Rate* in eq. (7) with respect to the case of sampling with replacement.

3.4.1 SRSWOR vs SRSWR. Simple Random Sampling without Replacement (SRSWOR) randomly samples $x_1, x_2, \ldots, x_k$ distinct items from a finite population of N elements. Differently from *Simple Random Sampling with Replacement (SRSWR)* where the same item may appear more than once in the sample and items are independent from each other, in SRSWOR, items in the sample are not independent from each other. As a consequence, the covariance between two items x_i and x_j, $i \neq j$, is not zero but it is proportional to the population variance σ^2, namely $\mathrm{Cov}(x_i, x_j) = -\frac{\sigma^2}{N-1}$ [10, 26]. It should be noted as this covariance is a fixed factor that depends only on the population size N and not on the sample size k.

The sample mean $\bar{X} = \frac{1}{k}\sum_{i=1}^{k} x_i$ is an unbiased estimator of the population mean for both SRSWOR and SRSWR. The variance of the sample mean in SRSWOR is $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{k}\frac{N-k}{N-1}$, where $\frac{\sigma^2}{k}$ is the same as the variance of the sample mean in SRSWR and $\frac{N-k}{N-1}$ is the *Finite Population Correction (FPC)* factor. In other words, the variance of the sample mean in the SRSWOR case is the variance of the sample mean in the SRSWR case adjusted by the FPC factor. When the population size N is large or the sample size k is small compared to N, $FPC \to 1$ and the variance of the sample mean is (almost) identical to the case of SRSWR. When the sample size k increases with respect to N, $FPC < 1$, the variance of the sample mean is reduced, making the sample mean more accurate and smaller confidence intervals could be considered around it [26]. If one does not consider the FPC and instead just use the variance of the sample mean as in the SRSWR case, the actual variance of the sample mean may be overestimated. For this reason, SRSWOR is said to be more *efficient* than SRSWR.

The sample variance $s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{X})^2$ is an unbiased estimator of the population variance in the case of SRSWR while in the SRSWOR case is no more an unbiased estimator of the population variance and it needs to be adjusted by the $\frac{N-1}{N}$ factor. The SRSWR sample variance tends to slightly overestimate the population variance in the SRSWOR case; as soon as the population size increases $\frac{N-1}{N} \longrightarrow 1$ and it becomes (almost) identical to the SRSWR case. The adjustment $\frac{N-1}{N}$ to the sample variance depends only on the population size N and not on the sample size k.

3.4.2 SRSWOR vs SRSWR for Topic Splitting. The bigger efficiency of SRSWOR with respect to SRSWR would give us the possibility of creating smaller confidence intervals around mean run performance and being more accurate in comparing two runs thus, possibly, obtaining more significantly different run pairs.

However, our set of topics T is just a *frame* of an "idealized" larger population of topics. In such cases, if one is interested in the "idealized" population rather than the frame, as suggested by Siegel [34], it is preferable to not apply the FPC factor and to proceed with the estimates as in the case of sampling with replacement. Moreover, we are using two topic sets TSs, calculating significance. The size of the TSs is the same and, therefore, the FPC factor or other biases, such as the adjustment in the sample variance estimation[2], would be the same for both sets. Since we are interested in a comparison across TSs rather than in absolute numbers, applying the same scaling on both sides would be less relevant.

Therefore, we use topic splitting based on SRSWOR, since it better represents the case of a set of topics used in an experiment and a different set of topics used in another experiment or in operation, or the case of two samples from a large query log of a search engine which are unlikely to overlap, or the case of topics changing over time, e.g. when you sample from a query log to run experiments and, meanwhile, users change their interests and start submitting other topics. In other terms, topic splitting based on SRSWOR allow us to better study the *generalizabilty* of the drawn inferences.

Sanderson and Zobel highlighted possible inflation of DR as the topic set size approaches $\frac{T}{2}$, we note that the covariance between items in the sample is fixed, just depending on the population size T and not on the sample size $ts_1 = ts_2$. As a consequence, a topic set size of $\frac{T}{2}$ does not cause more dependence in the items of the sample than any other size. Therefore, if SRSWOR should have any impact on the DR with respect to SRSWR, this should be more or less the same for all the topic set sizes. Moreover, the covariance $\mathrm{Cov}(x_i, x_j) = -\frac{\sigma^2}{T-1}$, $i \neq j$ is inversely proportional to the total number of topics T which we are using, which usually is in the order of tens or hundreds, and this makes it a small factor, further reducing the difference between SRSWOR and SRSWR.

4 EXPERIMENTAL SETUP

Collections. We used the following collections: TREC 13 (T13) robust track [39], which contains 249 topics, 110 runs; TREC 26 (T26) Common Core track [1], which contains 50 topics, 75 runs; TREC 27 (T27) Common Core track, which contains 50 topics, 72 runs.

We used Average Precision [7], Precision at 10, and *Normalized Discounted Cumulated Gain (nDCG)* [19]. For binary measures, the multi-graded judgements were mapped to binary: everything above not relevant was considered relevant. Finding little difference between the three measures, we report nDCG only.

Topic Sets. We sampled topics forming two TSs. For T13, we formed splits of 2%, 4%, 10%, 20%, and 50%, containing, respectively, 5, 10, 25, 50, and 125 topics each[3]. For T26 and T27, we formed splits of 10%, 20%, and 50% containing, respectively, 5, 10, and 25 topics each. For each split, we repeated the topic sampling $S = 1,000$ times. For the run set-wise count of Section 3.2, we took the arithmetic mean, resulting in counts having non-integer values.

We performed the sampling both without replacement (WOR), the main experiments in the paper, and with replacement (WR) to compare the two alternatives as discussed in Section 3.4. When it is not explicitly mentioned, we are reporting WOR results.

Significance Level and Multiple Comparison. We set the significance level $\alpha = 0.05$. In order to control for the increased Type-I errors due to the multiple comparisons between all possible pairs, we adopt the Tukey HSD correction [17, 37] for the anv2 ANOVA model and Bonferroni's correction [5] for the ttpB paired Student's t test. We also run a Student's t test with No Bonferroni correction (ttpNB), to investigate the impact of this practice in our field.

Reproducibility. Code at: https://bitbucket.org/frrncl/sigir2024-fs/.

5 ANALYSIS

5.1 Multiple Testing Correction

We start by comparing significance tests with and without multiple corrections. As detailed in Sec 3.1 the application of correction methods, such as Bonferroni or Tukey HSD, reduces the number of Type I errors by raising the p-value threshold. Consequently, for this analysis, we count the number of significant pairs across the two topic sets, TS1 and TS2.

We first measure on the largest possible split of topic sets in the three collections, T13, T26, and T27 measuring on both conditions: track and participant. For the participant condition, we aggregate the pairs of the 13 participant groups in T13, 14 groups in T26, and 11 groups in T27. We can see in Table 1 the number of significant pairs from ttpNB is substantially higher than anv2 and ttpB in both the track and participant conditions. We see in the TS1% and TS2% columns that the multiple comparison correction of anv2 and ttpB leads to fewer significant pairs: 39% – 60% of the number of pairs found in TTPNB. Examining the track and participant conditions, the fraction of significant pairs resulting from multiple testing correction is broadly similar.

The larger number of topics in the T13 collection allows an examination of the impact of multiple testing correction across topic set size splits, see Table 2. For smaller topic set sizes (5, 10, 20),

[2] which depends on T and not on the topic set size

[3] As there are 249 topics in T13, the 50%-50% split is actually 125-124.

the impact of multiple testing in reducing the number of significant pairs is greater for the track condition compared to the participant condition. As the topic set sizes grow (100, 125), the difference between the two conditions reduces and becomes broadly similar for the larger topic set sizes. We also find that there is a notable difference between the correction from Bonferroni compared to Tukey HSD. For the small topic sizes (5, 10) there are virtually no significant results after the Bonferroni correction. However, for the Tukey HSD test, while the number of significant tests is reduced substantially, there is still a notable number found to be significant.

The application of multiple testing correction is a proven method for removing false positives from experimental results. Focusing on topic set sizes that might be used in offline testing (25 topics or more), the reduction in number of significance tests resulting from the multiple testing correction ranges from a low 32% to a high of 67%. The implication from this pair of measurements is that uncorrected significance tests, common in a great many publications, are routinely reporting false positive results.

5.2 Track vs Participant Conditions

Using topic splitting, we examine if a significant result found in one split is repeated in another. Bias, eq. (6), measures the fraction of significant differences in an experiment that run the risk of being incorrect. We contrast bias measured in the two conditions: track and participant. As we see on the left of Figure 1 the bias in the participant only runs is substantially higher than for track runs.

Examining p_{Bias} (graphs on the right of Figure 1), we observe that the probability of a pair contributing to bias is higher in the participant condition than in the track, reinforcing previous observations about the systematic difference between the conditions.

The decrease in *bias* as the topic set size increases is expected: more topics should mean more stable measurement. The somewhat increasing trend of p_{Bias} might look surprising. However, as shown in Table 2, smaller topic set sizes lead to a much smaller overall number of significantly different pairs; therefore, each pair has less probability to be significantly different and, in turn, less probability to be an AA, MA or MD. Therefore, while increasing the number of topics is beneficial to get better estimates and lower the overall amount of bias (Figure 1 on the left), at the same time, it also slightly increases the probability of a pair to contribute to the smaller amount of bias left (Figure 1 on the right).

Obtaining significance increases the likelihood of a paper being published [11]. We contend that the participant condition is a more realistic simulation of the situation researchers find themselves when conducting and potentially publishing research. They will be testing the difference between retrieval systems that are producing similar scores with only minor differences between those scores. If significance is obtained, then the likelihood of publication is increased however, if as shown in the bias measures on the left, a large fraction of those significance measures are only inconsistently measured across different topic sets, then there is a substantial risk that the significance result is a false positive.

5.3 Fine-grained Run-Pairwise Analysis

Figure 2 shows an example of how to use the run pair-wise probabilities discussed in Section 3.3 to conduct a finer-grained analysis for participant RMIT [3] in track T27 using nDCG, according to different significance tests – anv2, ttpB, and ttpNB, one for each subfigure. We examine the five submitted runs of RMIT for T27 resulting in ten runs pairs to be compared. One can use this analysis as a tool to mitigate the inconsistencies discussed in the previous sections and to support researchers in taking more principled decisions.

The y-axis reports each pair of runs for a participant; next to the label, in parentheses, we report the mean nDCG for that run using all the topics of T27; the pair is in bold when it is significantly different according to the significance test conducted using all the topics of T27; labels are ordered from the best performing run at the bottom of the figure to the worst performing at the top. In short, the y-axis labels summarize the usual setup, i.e. what a participant sees about their runs when using all the data of a track.

For each pair of runs the figure shows a horizontal stacked bar chart where each segment corresponds to one of the probabilities p_{AA} (in green), p_{PA} (in light gray), p_{PD} (in dark gray), p_{MA} (in yellow), p_{MD} (in orange), and p_{AD} (in red), ordered left-to-right by their desirability/severity. Since the probabilities sum up to 1, the x-axis goes from 0 to 1. Finally, a violet diamond indicates p_{Bias} while a red circle indicates p_{DR}. The probabilities are estimated using the 50%-50% topic split (25 topics in each), sampling without replacement, and repeating the sampling S times.

We present some (though not all) of the examinations possible with Figure 2. According to anv2 in Figure 2a, the top run (RMITUQVDBFNZDM1, nDCG = 0.7190) is not significantly different from the second-top run (RMITUQVDBFDM3, nDCG = 0.7143) while both of them are significantly different from the third-top run (RMITFDA4, nDCG = 0.6525). However, the pair (RMITUQVDBFNZDM1, RMITFDA4) has a probability of mixed agreement $p_{MA} = 0.76$ which leads to a probability of originating bias $p_{Bias} = 0.76$ while the pair (RMITUQVDBFDM3, RMITFDA4) has a lower probability of mixed agreement $p_{MA} = 0.46$ which leads to a much smaller probability of originating bias $p_{Bias} = 0.46$. Therefore, if RMITFDA4 was the IR system currently in production and RMITUQVDBFNZDM1 and RMITUQVDBFDM3 were two alternative improved versions, it could be preferable to go for RMITUQVDBFDM3 instead of RMITUQVDBFNZDM1 since they are both significantly better than RMITFDA4 and not significantly better than each other but RMITUQVDBFDM3 promises to deliver more consistent and predictable improvements.

Suppose that the run (RMITUQVBestDM2, nDCG = 0.6410) is the system in production. The pair (RMITUQVDBFNZDM1, RMITUQVBestDM2) has a probability of active agreement $p_{AA} = 0.38$ and of mixed agreement $p_{MA} = 0.62$ which leads to a probability of originating bias $p_{Bias} = 0.62$; on the other hand, the pair (RMITUQVDBFDM3, RMITUQVBestDM2) has a probability of active agreement $p_{AA} \sim= 0.08$ and of mixed agreement $p_{MA} = 0.92$ which leads to a probability of originating bias $p_{Bias} = 0.92$. Therefore, it would be preferable to put in production RMITUQVDBFNZDM1 instead of RMITUQVDBFDM3 thanks to the lower probability of bias and high probability of active agreements, which give grounds for expecting consistent and predictable improvements.

We reach similar conclusions for ttpB in Figure 2b, even if the Student's t test with Bonferroni's correction tends to be more "optimistic" in detecting significance compared to anv2. For example, for the pair (RMITUQVDBFNZDM1, RMITFDA4), there is a probability of active agreement $p_{AA} = 0.03$ (while with anv2 it was $p_{AA} = 0.00$)

Table 1: Comparing with and without multiple comparisons across collections

		Pairs	Sig. Pairs				Pairs	Sig. Pairs			
Collection	Model	(track)	TS1 avg.	TS2 avg.	TS1%	TS2%	(participant)	TS1 avg.	TS2 avg.	TS1%	TS2%
T13 (125, 124)	TTPNB	5995	4721.2	4714.5			435	264.3	263.5		
	ANV2		3134.3	3118.5	66%	66%		166.6	165.5	63%	63%
	TTPB		3158.7	3144.6	67%	67%		178.1	176.8	67%	67%
T26 (25, 25)	TTPNB	2775	1926.5	1917.6			254	122.5	122.7		
	ANV2		1120.7	1106.4	58%	58%		58.0	57.8	47%	47%
	TTPB		781.9	776.3	41%	40%		48.0	48.2	39%	39%
T27 (25, 25)	TTPNB	2556	1679.5	1688.3			249	86.6	87.5		
	ANV2		995.1	999.1	59%	59%		54.2	54.7	63%	63%
	TTPB		804.7	806.5	48%	48%		41.8	41.8	48%	48%

Table 2: Comparing with and without multiple comparisons across different topic set sizes

T13 Topic Set		Significant Pairs (track, 5995)				Significant Pairs (participant, 435)			
(TS1, TS2) size	Model	TS1 avg.	TS2 avg.	TS1%	TS2%	TS1 avg.	TS2 avg.	TS1%	TS2%
wor_02_02 (05, 05)	TTPNB	1254.3	1245.3			31.2	30.4		
	ANV2	327.1	330.1	26%	27%	19.7	19.6	63%	65%
	TTPB	0.5	0.5	0%	0%	0.9	0.9	3%	3%
wor_04_04 (10, 10)	TTPNB	2223.1	2246.8			71.9	73.0		
	ANV2	701.4	716.0	32%	32%	35.1	35.8	49%	49%
	TTPB	23.8	25.7	1%	1%	6.0	6.5	8%	9%
wor_10_10 (25, 25)	TTPNB	3361.4	3360.4			142.9	142.7		
	ANV2	1431.6	1429.8	43%	43%	72.2	71.5	51%	50%
	TTPB	888.5	883.5	26%	26%	44.9	44.1	31%	31%
wor_20_20 (50, 50)	TTPNB	4041.5	4036.2			197.8	198.5		
	ANV2	2103.6	2097.6	52%	52%	108.8	108.5	55%	55%
	TTPB	1868.6	1862.9	46%	46%	99.4	99.3	50%	50%
wor_30_30 (75, 75)	TTPNB	4373.3	4365.7			229.1	228.3		
	ANV2	2557.8	2546.2	58%	58%	133.8	133.4	58%	58%
	TTPB	2458.9	2445.8	56%	56%	132.7	132.2	58%	58%
wor_40_40 (100, 100)	TTPNB	4577.4	4577.3			249.6	249.8		
	ANV2	2881.1	2886.0	63%	63%	152.0	152.0	61%	61%
	TTPB	2860.7	2862.0	62%	63%	157.4	157.3	63%	63%
wor_50_50 (125, 124)	TTPNB	4721.2	4714.5			264.3	263.5		
	ANV2	3134.3	3118.5	66%	66%	166.6	165.5	63%	63%
	TTPB	3158.7	3144.6	67%	67%	178.1	176.8	67%	67%

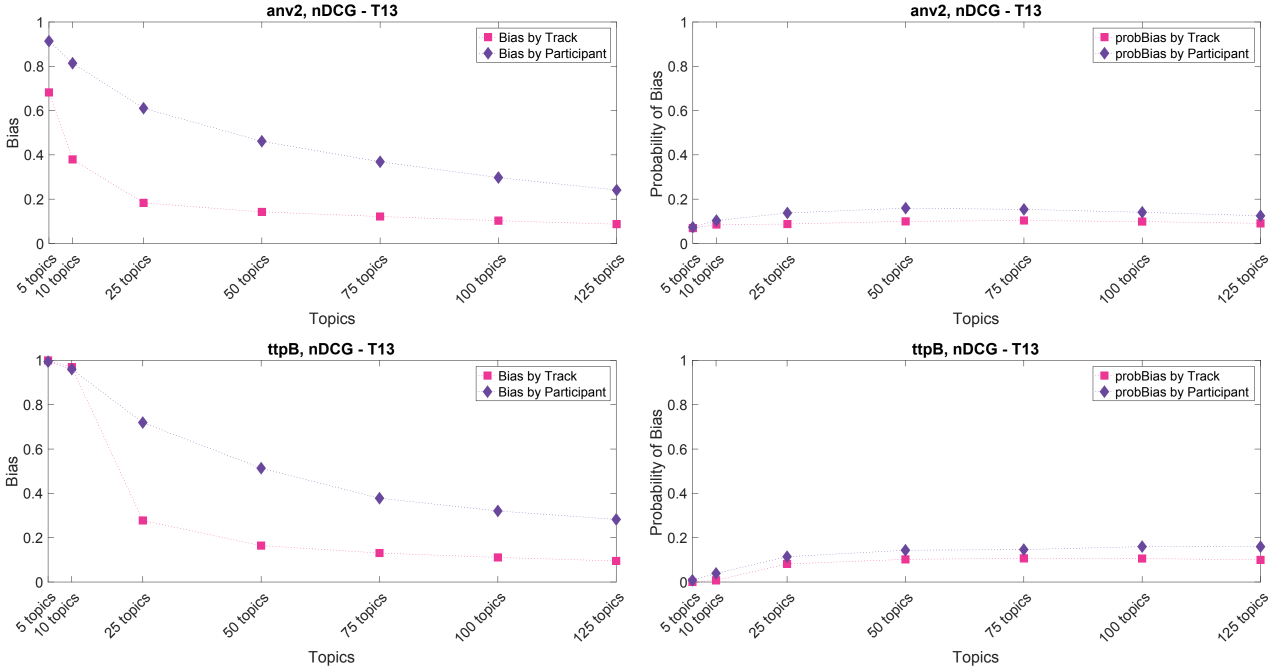
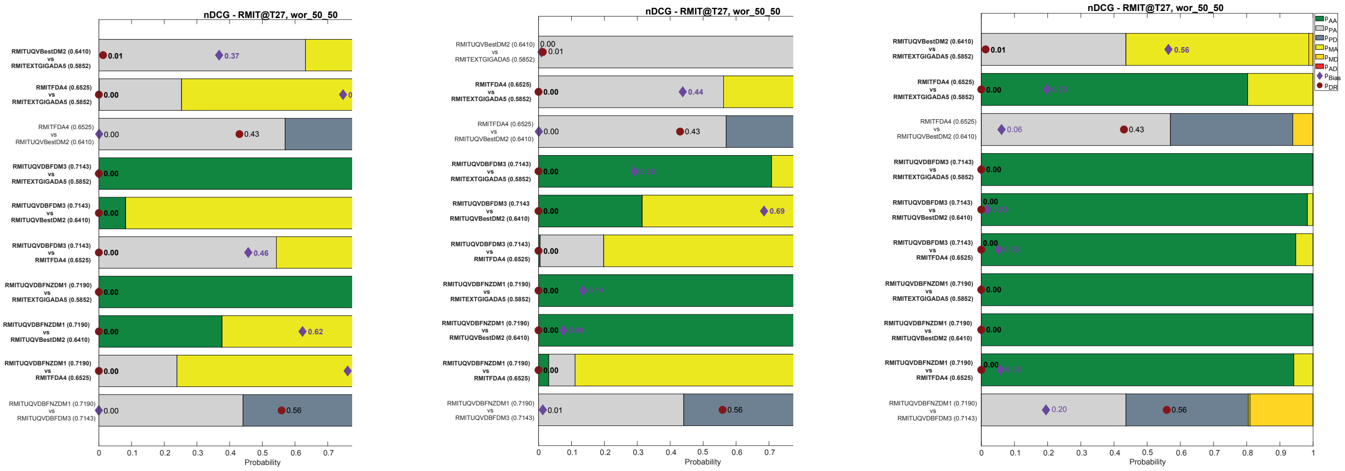


Figure 1: Track vs Participant condition on T13 for nDCG using anv2 (top) and ttpB (bottom).



(a) Participant RMIT in track T27 usir

(b) Participant RMIT in track T27 usi

(c) Participant RMIT in track T27 using ttpNB.

Figure 2: Pair-wise probability analysis of significance tests for participant RMIT in T27 using nDCG.

and the probability of mixed agreement goes to $p_{MA} = 0.89$ (while with anv2 it was $p_{MA} = 0.76$) originating a probability of bias $p_{Bias} = 0.89$ (while with anv2 it was $p_{Bias} = 0.76$). Besides the difference in the normalization factor for the mean performance differences, the main reason of the difference between anv2 and ttpB lays in their thresholds: $\frac{1}{\sqrt{25}} t_{24}^{1-\frac{0.05/10}{2}} = 0.6181$ is more liberal than $\frac{1}{\sqrt{25}} q_{5,96}^{0.05} = 0.7864$.

Finally, Figure 2c shows multiple testing without any correction. Ignoring FWER causes the probability of active agreements p_{AA} to grow for almost all the pairs of runs, incorporating previously mixed agreements. Moreover, the rise in significant differences causes an increase in the probability of mixed and active disagreements, as in the case of the pair (RMITFDA4, RMITUQVBestDM2) which now has $p_{AD} = 0.001$. Therefore, even if we generally consider higher p_{AA} and lower p_{MA} better, if in our plot we observe very high and predominant green bars (high p_{AA}) at a price of a substantial decrease in the yellow bars (low p_{MA}), possibly accompanied by the appearance of some orange bars (increase in p_{MD}) or even red ones (increase in p_{AD}), we should take this as an indicator of possible serious issues in our experimental setup, as for example a lack of adjustment for multiple comparisons.

5.4 Topic Splitting With/Without Replacement

Investigating replacement, see Section 3.4, Tables 3 and 4 report *DR* and *Bias* for different topics set sizes for nDCG when splitting topics using sampling without replacement (WOR) and with replacement (WR). As *DR* is independent from the statistical significance test at hand, the tables do not indicate a significance test for it. *Bias* depends on the significance test and, for space reasons, we report it only for anv2.

Differently from what was highlighted by Sanderson and Zobel [33], we observe the same behaviour and almost the same *DR* values for both WR and WOR, in both the track and participant conditions. Moreover, when the topic set size is $\frac{T}{2}$, i.e. 125 T13 and 25 T26 and T27, we do not observe any inflation of the *DR* for the sampling without replacement case. We reach the same conclusions for *Bias*. Therefore, we conclude that the sampling without replacement is not introducing specific deviations and that we can adopt it for the reasons discussed in Section 3.4.

Tables 3 and 4 provide further evidence about the difference between the track and participant conditions, discussed in Section 5.2, showing how *Bias*, but often also *DR*, can be substantially worse for the participant condition than for track. They provide a more fine-grained view than Figure 1, since they report an example for a given participant, together with confidence intervals around it, instead of just an aggregation across all the participants' groups; they also show how this difference holds for both WR and WOR.

6 DISCUSSION

We reflect on the results we obtained from our experiments.

6.1 Correcting Multiple Testing

The results from Section 5.1 showed that 40% − 50% of uncorrected significance tests in an offline evaluation setting are likely Type I errors. As a community, we need to adopt multiple testing correction for publications. Past work [4, 6] empirically showed the importance of multiple test corrections to limit the potential of Type I errors occurring but that work did not quantify errors in the manner detailed here. Later others [16, 31] outlined the statistical argument for using that form of correction, but did not estimate how many Type I errors might occur. Our experiments add to the body of knowledge by showing the magnitude of those errors when testing on a set of pair-wise comparisons across a series of runs: a typical IR experiment.

Our results engage with statements made by Carterette [8] who stated that the correction in p-value for *participant* runs (Option 2a in his paper) would be the smallest that one would see across different multiple testing scenarios. One might imagine that this smaller correction might lead to a smaller reduction in the number of significance tests found in our experiments, however no such

Table 3: *DR* **and** *Bias* **for nDCG using topic split sampling with (WR) or without (WOR) replacement for on T13.**

Condition	Measure		5 Topics	10 Topics	25 Topics	50 Topics	75 Topics	100 Topics	125 Topics
T13	*DR*	WOR	0.2510 ± 0.0033	0.1986 ± 0.0026	0.1329 ± 0.0017	0.0965 ± 0.0013	0.0791 ± 0.0010	0.0684 ± 0.0009	0.0608 ± 0.0007
		WR	0.2511 ± 0.0034	0.1929 ± 0.0027	0.1293 ± 0.0017	0.0951 ± 0.0013	0.0774 ± 0.0010	0.0677 ± 0.0009	0.0610 ± 0.0008
FUB@T13	*DR*	WOR	0.3912 ± 0.0101	0.3347 ± 0.0092	0.2176 ± 0.0065	0.1500 ± 0.0047	0.1212 ± 0.0036	0.1070 ± 0.0033	0.0914 ± 0.0026
		WR	0.3799 ± 0.0104	0.3243 ± 0.0090	0.2132 ± 0.0066	0.1459 ± 0.0047	0.1190 ± 0.0039	0.1026 ± 0.0034	0.0933 ± 0.0032
T13, anv2	*Bias*	WOR	0.6821 ± 0.0116	0.3796 ± 0.0082	0.1834 ± 0.0028	0.1423 ± 0.0018	0.1217 ± 0.0015	0.1026 ± 0.0013	0.0871 ± 0.0011
		WR	0.6786 ± 0.0116	0.3733 ± 0.0083	0.1809 ± 0.0030	0.1412 ± 0.0018	0.1181 ± 0.0015	0.1010 ± 0.0014	0.0883 ± 0.0012
FUB@T13, anv2	*Bias*	WOR	0.9835 ± 0.0122	0.9835 ± 0.0089	0.9294 ± 0.0120	0.7075 ± 0.0165	0.4794 ± 0.0129	0.3406 ± 0.0093	0.2241 ± 0.0073
		WR	0.9942 ± 0.0066	0.9884 ± 0.0071	0.8969 ± 0.0145	0.6578 ± 0.0172	0.4550 ± 0.0143	0.3356 ± 0.0109	0.2547 ± 0.0081

Table 4: Same measures as Table 3 based on T26 and T27.

Condition	Measure		5 Topics	10 Topics	25 Topics
T26	*DR*	WOR	0.1875 ± 0.0033	0.1373 ± 0.0020	0.0989 ± 0.0013
		WR	0.1795 ± 0.0034	0.1343 ± 0.0024	0.0892 ± 0.0014
SABIR@T26	*DR*	WOR	0.2884 ± 0.0070	0.2201 ± 0.0058	0.1299 ± 0.0039
		WR	0.2791 ± 0.0072	0.2113 ± 0.0057	0.1370 ± 0.0042
T27	*DR*	WOR	0.2002 ± 0.0030	0.1557 ± 0.0021	0.1224 ± 0.0013
		WR	0.1908 ± 0.0032	0.1505 ± 0.0023	0.1049 ± 0.0016
RMIT@T27	*DR*	WOR	0.1739 ± 0.0074	0.1208 ± 0.0053	0.1002 ± 0.0046
		WR	0.1718 ± 0.0078	0.1177 ± 0.0054	0.0913 ± 0.0044
T26, anv2	*Bias*	WOR	0.4073 ± 0.0089	0.2513 ± 0.0057	0.1885 ± 0.0033
		WR	0.3986 ± 0.0096	0.2540 ± 0.0059	0.1656 ± 0.0035
SABIR@T26, anv2	*Bias*	WOR	0.9783 ± 0.0079	0.9054 ± 0.0129	0.3092 ± 0.0089
		WR	0.9467 ± 0.0136	0.8321 ± 0.0166	0.3164 ± 0.0095
T27, anv2	*Bias*	WOR	0.2243 ± 0.0083	0.1495 ± 0.0027	0.0954 ± 0.0017
		WR	0.2279 ± 0.0087	0.1484 ± 0.0030	0.0964 ± 0.0020
RMIT@T27, anv2	*Bias*	WOR	0.9244 ± 0.0196	0.5383 ± 0.0249	0.4385 ± 0.0074
		WR	0.8738 ± 0.0229	0.5578 ± 0.0240	0.3048 ± 0.0104

difference between *track* and *participant* runs was observed. The TS1% and TS2% values are similar between *track* and *participant* in Tables 1 & 2. We view this lack of difference between the *participant* and *track* conditions to be an indication of the correction methods working appropriately. If one is conducting experiments with a large number of comparisons, as occurs in a *track*, the level of correction needs to be greater in order to avoid Type I errors. Note, we see that the differences between *track* and *participant* varies depending on the topic set size. Carterette [8] did not examine such an aspect of the problem.

6.2 Track vs Participant Conditions

The results in Section 5.2 contradict research results shown in past work. In Voorhees and Buckley [40] it was suggested that *participant* only runs would result in less error when comparing runs across different topics splits. We find the opposite of this: error in the *participant* condition is higher than the *track* condition. Reexamining Voorhees and Buckley's experiments, we wonder if this contradiction is due to the way that the *participant* only runs were used in that work. In Figure 5 of the past work, a comparison was made between *track* and *participant* conditions. However, it would appear that there is a difference in the way that runs are considered. While in the *track* condition the bottom 25% of runs were discarded (as normal), in the *participant* condition they were not. Moreover, in both conditions, Voorhees and Buckley aggregate data from different tracks and rearrange them by the absolute performance delta in the pair; however, the same absolute performance delta may be small or big depending on the track and, consequently, more or less prone to swapping. We speculate that this difference in the range of runs used and the way in which data points were aggregated explains the apparent contradiction.

6.3 Fine-grained Run Pairwise Analysis

We have shown how multiple comparisons and working in track or participant conditions can lead to different types of error. We have also shown how overall approaches, i.e. properly adjusting for multiple comparisons to control the Type I error rate or increasing the topic set size for reducing the publication bias, make experimental conclusions more reliable. If we find that run pairs are significantly different, how do we decide which are the best? Up to know, the only criterion was the higher performance, the better; however, we have seen that inconsistencies may be hidden in almost all the performance ranges. Therefore, our answer to address this issue was to introduce the probabilities of p_{AA}, p_{AD}, . . ., p_{Bias}, and p_{DR} to be able to conduct a more fine-grained analysis on each single pair and take a more informed decision based on how much more reliable that pair is likely to be.

7 CONCLUSIONS

We asked if there is a prevalence of Type I errors in IR experiments of the type typically published in IR forums? The answer appears to be yes. Experimenting across three test collections, we established that without the use of multiple testing correction, a large number (40% – 50%) of statistically significant results in IR experiments are likely to be Type I errors. Then we questioned if the error varies when measured on the runs of single participants vs. all the runs submitted to an evaluation track? Here we found no difference. However, for tests run with different topics sizes, the number of Type I errors increased as topic sizes reduced.

We ran experiments showing that there is a high level of inconsistency in significant retrieval results conducted on participant only runs, the family of runs that one typically sees in published research. The level of inconsistency was 2-3 times higher for participant only compared to experiments using all runs from a track.

Combining both of these results we are forced to conclude that the level of Type I error in IR experiments using offline testing resources is likely to be extremely high.

Given such inconsistency, how best can it be alleviated? The most effective way of alleviating many of these Type I errors is to implement correction through multiple testing as a standard in IR publications. We also detailed an analysis of IR experiments that allows an experimenter to understand whether a result is more likely to be a Type I error or a genuine result.

Acknowledgments. This research was conducted in part by the ARC Centre of Excellence for Automated Decision-Making and Society, and funded partially by the Australian Government through the Australian Research Council (CE200100005). This research was partially supported by CAMEO, PRIN 2022 n. 2022ZLL7MW.

REFERENCES

[1] J. Allan, D. K. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees. 2018. TREC 2017 Common Core Track Overview. In *The Twenty-Sixth Text REtrieval Conference Proceedings (TREC 2017)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-324, Washington, USA.

[2] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval* 1, 1-2 (May 1999), 7–34.

[3] R. Benham, L. Gallagher, J. Mackenzie, B. Liu, X. Lu, F. Scholer, A. Moffat, and J. S. Culpepper. 2019. RMIT at the 2018 TREC CORE Track. In *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-331, Washington, USA.

[4] Roi Blanco and Hugo Zaragoza. 2011. *Beware of relatively large but meaningless improvements*. Technical Report. Yahoo! Research 2011–001.

[5] C. E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Number 8 in Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze. Libreria internazionale Seeber, Firenze, Italia.

[6] L. Boytsov, A. Belova, and P. Westfall. 2013. Deciding on an Adjustment for Multiplicity in IR Experiments, See [20], 403–412.

[7] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*, D. K. Harman and E. M. Voorhees (Eds.). MIT Press, Cambridge (MA), USA, 53–78.

[8] B. A. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 4:1–4:34.

[9] B. A. Carterette. 2017. But Is It Statistically Significant? Statistical Significance in IR Research, 1995-2014. In *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White (Eds.). ACM Press, New York, USA, 1125–1128.

[10] W. G. Cochran. 1977. *Sampling Techniques* (3rd ed.). John Wiley & Sons, New York, USA.

[11] P. J. Easterbrook, R. Gopalan, J. A. Berlin, and D. R. Matthews. 1991. Publication bias in clinical research. *The Lancet* 337, 8746 (1991), 867–872.

[12] N. Ferro and M. Sanderson. 2019. Improving the Accuracy of System Performance Estimation by Using Shards. In *Proc. 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer (Eds.). ACM Press, New York, USA, 805–814.

[13] N. Ferro and M. Sanderson. 2022. How do you Test a Test? A Multifaceted Examination of Significance Tests. In *Proc. 15th ACM International Conference on Web Searching and Data Mining (WSDM 2022)*, K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, and J. Tang (Eds.). ACM Press, New York, USA, 280–288.

[14] N. Ferro and G. Silvello. 2016. A General Linear Mixed Models Approach to Study System Component Effects, See [25], 25–34.

[15] R. A. Fisher. 1925. *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, UK.

[16] N. Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (December 2017), 32–41.

[17] Y. Hochberg and A. C. Tamhane. 1987. *Multiple Comparison Procedures*. John Wiley & Sons, USA.

[18] J. C. Hsu. 1996. *Multiple Comparisons. Theory and methods*. Chapman and Hall/CRC, USA.

[19] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (October 2002), 422–446.

[20] G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai (Eds.). 2013. *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*. ACM Press, New York, USA.

[21] S. E. Maxwell, H. D. Delaney, and K. Kelly. 2017. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge, Taylor & Francis Group, UK.

[22] D. Newman. 1939. The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. *Biometrika* 31, 2 (July 1939), 20–30.

[23] J. Parapar, D. E Losada, and Á. Barreiro. 2021. Testing the tests: simulation of rankings to compare statistical significance tests in information retrieval evaluation. In *Proc. 36th ACM/SIGAPP Symposium On Applied Computing (SAC 2021)*, C.-H. Hung, J. Hong, A. Bechini, and E. Song (Eds.). ACM Press, New York, USA, 655–664.

[24] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, and A. Barreiro. 2020. Using Score Distributions to Compare Statistical Significance Tests for Information Retrieval Evaluation. *Journal of the American Society for Information Science and Technology (JASIST)* 71, 1 (January 2020), 98–113.

[25] R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel (Eds.). 2016. *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM Press, New York, USA.

[26] J. A. Rice. 2007. *Mathematical Statistics and Data Analysis* (3rd ed.). Thomson, Belmont (CA), USA.

[27] S. E. Robertson and E. Kanoulas. 2012. On Per-topic Variance in IR Evaluation. In *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, W. Hersh, J. Callan, Y. Maarek, and M. Sanderson (Eds.). ACM Press, New York, USA, 891–900.

[28] A. Rutherford. 2011. *ANOVA and ANCOVA. A GLM Approach* (2nd ed.). John Wiley & Sons, New York, USA.

[29] T. Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015, See [25], 5–14.

[30] T. Sakai. 2018. *Laboratory Experiments in Information Retrieval*. The Information Retrieval Series, Vol. 40. Springer Singapore.

[31] T. Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1 (June 2020), p14:1–p14:8.

[32] M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)* 4, 4 (2010), 247–375.

[33] M. Sanderson and J. Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). ACM Press, New York, USA, 162–169.

[34] A. F. Siegel. 2016. *Practical Business Statistics* (7th ed.). Elsevier, The Netherlands.

[35] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (March 1908), 1–25.

[36] J. M. Tague-Sutcliffe and J. Blustein. 1995. A Statistical Analysis of the TREC-3 Data. In *The Third Text REtrieval Conference (TREC-3)*, D. K. Harman (Ed.). National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 385–398.

[37] J. W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (June 1949), 99–114.

[38] J. Urbano, M. Marrero, and D. Martín. 2013. A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation, See [20], 925–928.

[39] E. M. Voorhees. 2005. Overview of the TREC 2004 Robust Track. In *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA.

[40] E. M. Voorhees and C. Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. Hyon Myaeng (Eds.). ACM Press, New York, USA, 316–323.

[41] E. M. Voorhees, D. Samarov, and I. Soboroff. 2017. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)* 36, 2 (September 2017), 12:1–12:21.

[42] F. Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (December 1945), 80–83.

[43] J. Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (Eds.). ACM Press, New York, USA, 307–314.