Statistical Power in Retrieval Experimentation

William Webber
Computer Science and
Software Engineering
The University of Melbourne
Victoria 3010, Australia
wew@csse.unimelb.edu.au

Alistair Moffat
Computer Science and
Software Engineering
The University of Melbourne
Victoria 3010, Australia
alistair@csse.unimelb.edu.au

Justin Zobel
NICTA VRL
The University of Melbourne
Victoria 3010, Australia
jz@csse.unimelb.edu.au

ABSTRACT

The power of a statistical test specifies the sample size required to reliably detect a given true effect. In IR evaluation, the power corresponds to the number of topics that are likely to be sufficient to detect a certain degree of superiority of one system over another. To predict the power of a test, one must estimate the variability of the population being sampled from; here, of between-system score deltas. This paper demonstrates that basing such an estimation either on previous experience or on trial experiments leaves wide margins of error. Iteratively adding more topics to the test set until power is achieved is more efficient; however, we show that it leads to a bias in favour of finding both power and significance. A hybrid methodology is proposed, and the reporting requirements of the experimenter using this methodology are laid out. We also demonstrate that greater statistical power is achieved for the same relevance assessment effort by evaluating a large number of topics shallowly than a small number deeply.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

General Terms

Experimentation, measurement.

Keywords

Retrieval experiment, evaluation, system measurement.

1. INTRODUCTION

The need to verify the statistical significance of results in information retrieval (IR) experiments is well established within the IR research community. Such statistical tests determine whether the observed difference in performance between two IR systems is significant, or whether it could have occurred by chance. However, if the experiment fails to find significance, then one cannot simply conclude that no consequential difference exists. Instead, the experimenter wishes to know how large an actual difference in performance could have been missed. Additionally, when designing an IR experiment, the experimenter needs to decide how large a test set is required to reliably detect the difference in performance that the experimenter would regard as being consequential.

How small a real difference a statistical significance test can reliably detect is referred to as the statistical *power* of that test. The IR experimenter is faced with the question of power when considering either using an existing test collection, such as one of those produced by the TREC effort [Voorhees and Harman, 2005], or purpose-building a new one. If a test collection of suitable content already exists, the experimenter needs to determine whether the collection contains enough topics for them to be confident of detecting a consequential true difference between the systems, a determination which can be guided by examination of the variability of previous experimental results on that collection. If no suitable test collection exists, or the existing ones do not contain enough topics, then the experimenter needs to decide how many new topics to include in the experiment to attain the desired power. Failure to achieve the required experimental power can result in an unproductive experiment, one from which neither a positive nor a negative conclusion can be drawn.

We examine three different methods of estimating the power of a proposed experiment during design phase: (1) with reference to past experience of similar experiments; (2) by performing a trial experiment then a main one; and (3) by an iterative approach. We demonstrate that the former two methods leave quite wide bounds on power estimates, requiring the experimenter to deploy almost twice as many topics in the design as will on average prove necessary in execution. We also demonstrate that the iterative method, while efficient, leads to a subtle bias towards overestimating both power and statistical significance; the degree of this bias is empirically estimated. The assessment-rich or fastidious researcher will therefore wish to avoid iterative estimation. However, for researchers who are assessment-scarce, we propose an experimental methodology based on a hybrid of the three methods, including iteration. We also specify the details the researcher needs to declare when reporting significance results arising from this methodology.

Finally, employing power analysis as our tool, we examine the question of whether it is better, in a pairwise experiment with new topics, to evaluate fewer topics deeply, or more topics shallowly. Our results strongly indicate that the shallower evaluation of more topics produces far greater experimental power than deep evaluation of fewer topics for the same relevance assessment effort.

2. PREVIOUS WORK

The first author to empirically assess the usefulness of statistical significance tests in IR research was Zobel [1998]. He demon-

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

strates that significance findings on one half of a topic set are highly likely to be confirmed by a same-signed difference in means on the other half. For the t test at significance level 0.05, the rate of confirmation is 0.97–0.98. Zobel finds the t test to be more reliable than the Wilcoxon and sign tests.

The Wilcoxon and sign tests are approximate non-parametric significance tests, where precise score deltas are replaced by ranks and signs respectively for computational simplicity. The growing computational capacity of computers enables instead the employment of magnitude-aware nonparametric tests based on randomized resampling. Savoy [1997] proposes the use of one such test in IR research, evaluating the bootstrap on a handful of sample systems. Sakai [2006] was the first to employ the bootstrap on a wide range of systems, for the purpose of determining which evaluation metrics are more likely to lead to statistical significance. Smucker et al. [2007] propose instead that the randomized permutation test should be used, due to the minimal assumptions it imposes. Smucker et al. compare the sign, Wilcoxon, t, bootstrap, and randomized permutation tests, finding the first two unreliable, and the latter three to give similar results in practice.

Voorhees and Buckley [2002] propose a measure called the *error rate*, which is the likelihood that finding system a better than system b on one randomly selected set of queries would be reversed on another randomly selected set. Voorhees and Buckley calculate the average error rate for different score deltas and topic set sizes across historical TREC runs. The aim is to be able to conclude that a mean AP delta of (say) 0.06 is 90% reliable on 50 topics. Unlike statistical power analysis, this approach does not account for score delta variability, and requires the assumption that the results of previous TREC runs are applicable to new systems and collections. Sanderson and Zobel [2005] incorporate statistical significance alongside absolute deltas when calculating error rates.

Recently, interest has focused on methods of estimating metric scores for a run without exhaustive (to a depth) assessment of documents returned by that run. Aslam et al. [2006] propose an unequal sampling method for estimating AP and other metrics, where a document's probability of being judged is proportional to the weight its rank has in the metric and an estimate of its prior probability of relevance. Instead of estimating an absolute score for a run, Carterette [2007] presents a method for estimating the probability that one run has a positive score delta compared to another. The documents judged are those that are likely to have the greatest impact on these deltas, based on their probability of relevance. Carterette proposes an "aggregation of expert opinions" model to estimate each unjudged document's probability of relevance. Buckley and Voorhees [2004] examine the situation where an existing, but incomplete, test collection is employed to assess a new system, and propose a new metric called BPref to handle this situation. Yilmaz and Aslam [2006] and Sakai [2007] present methods for estimating existing metrics with incomplete judgments. Moffat and Zobel [2009] propose a metric, called rank-biased precision (RBP), in which the degree of uncertainty due to incomplete judgment is precisely quantified. Moffat et al. [2007] use this quantification to target judgments towards achieving greater certainty for the scores of well-performing systems. All such methods that attempt to estimate a metric introduce a new form of variability in run scores and score deltas which can be incorporated into a power analysis.

Cormack and Lynam [2007] define power as the probability that true significance will be achieved, without specification of a hypothesised true δ. Using this definition, they empirically determine the power of the t, Wilcoxon, and sign tests, finding the t test to be the most powerful (as well as the most reliable). Their definition of power is somewhat different from the classical one, and cannot be directly employed to measure the power of a test on a particular system pair, either during experimental design or post-hoc.

Carterette and Smucker [2007] examine statistical hypothesis testing in the presence of uncertainty about system scores on individual runs, particularly as it relates to the delta AP measure, described in Carterette [2007]. The significance test they deploy is the sign test. They provide an introduction to power analysis for this test. The detectable effect size is specified in terms of the proportion of topics one system outperforms the other on, not in terms of absolute metric difference This is similar to defining effect size in terms of the ratio of delta to standard deviation in paired t tests, as we consider in Section 3.3. With effect size so defined, problems of design-phase variability estimation fall away; however, this may not be the form in which the experimenter wishes to define effect. The sign test has been found by a number of researchers [Zobel, 1998, Smucker et al., 2007] to be the least reliable of the hypothesis tests. We instead examine the t test, and consider the case where the experimenter wishes to specify effect in absolute terms of the IR evaluation metric employed.

Carterette and Smucker consider the trade-off between assessing a smaller number of topics to certainty, or a larger number with a residual uncertainty as to true delta signedness, and conclude that shallow evaluation of more topics gives stronger power for the same effort, which we corroborate (under slightly different assumptions) for the t test in Section 6. Carterette et al. [2008] investigate the tradeoff between breadth and depth for score estimation on the TREC 2007 Million Query Track.

3. POWER AND EFFECT

3.1 Hypothesis testing

IR systems are evaluated and compared using a *test collection*, consisting of a document *corpus* and a set of *topics*. A topic is a user information need, often with an explicit description, which is formulated as a *query* and run by each IR system. A topic also includes human judgments as to which documents in the collection are *relevant* to that topic. These judgments may have been made prior to the experiment, or else they may be made after the experiment has been run, by judging the documents returned by the systems being evaluated. Let the set of n topics be $T = \{t_1, \cdots, t_n\}$.

Consider two IR systems, a and b, that are to be evaluated and compared using the test collection. Each IR system indexes the corpus, then runs the topics formulated as queries, and for each topic produces a ranked *run* of documents that it considers relevant for that topic. The documents are marked for relevance using the judgments for the topic, and a *metric* is used to produce a score for the run. Let the metric score that system a achieves on topic t be denoted as $m_{a,t}$. The aggregate or mean score $\overline{m_a}$ for system a is then $\sum_t m_{a,t}/n$, and similarly for system b. The difference between means, $\overline{m_a} - \overline{m_b}$, we denote as $d_{a,b}$ or simply d. It represents the *observed delta* between the systems. For each topic t, the per-topic delta $d_t = m_{a,t} - m_{b,t}$. Of course, $d = \sum_t d_t/n$.

Having observed $d_{a,b} > 0$ (informally, $a > b$), we can conclude that system a has outperformed system b on topic set T, at least under the metric employed. However, before rushing to print, we must verify that this observed improvement represents a real difference between the two systems, as it may have occurred by chance. We assume that the set of test topics T has been randomly sampled from the full population of topics $\mathcal{T}$, however that population is conceived. Therefore, the observed set of score differences $D = \{d_1, \cdots, d_n\}$ between systems a and b is randomly sampled from the population of score differences $\mathcal{D}$ between the two systems. The *true delta*, δ, between the systems is the mean of the population of deltas, $\delta = \bar{\mathcal{D}}$. Testing for *significance* involves formulating a *null hypothesis* H_0 that the two systems have in fact identical effectiveness, that is, that $\delta = 0$, and then determining the probability p that the observed difference d or greater could have occurred by chance if this hypothesis were true. If p is below some predetermined *significance level* α (where $\alpha = 0.05$ is a common choice), then H_0 is rejected, and the alternative hypothesis that the two systems are not equivalent is accepted.

Several hypothesis tests are available to the researcher; here we focus on the t test. The t test is applied by matching the observed outcome against quantiles of the t distribution, the sampling distribution of the mean of a normally distributed variable. If the sample size is large enough (greater than 25 is one conventional watershed), then the t test can be employed even if it is not known whether the underlying population is normally distributed, by invoking the Central Limit Theorem (CLT). In a one-tailed, paired t test on topic score deltas, where the number of topics is n, the mean delta is d, and the sample standard deviation of deltas is s, we calculate the t statistic:

$$t = \frac{d}{s/\sqrt{n-1}} \tag{1}$$

and check whether this statistic is greater than the $1 - \alpha$ quantile of the t distribution with $n - 1$ degrees of freedom.

Recent work [Sakai, 2006, Smucker et al., 2007] has suggested that the bootstrap or the randomized permutation test should be preferred to the t test due to the less stringent assumptions they make about the underlying distribution of the data. However, Smucker et al. [2007] find that the t, randomized permutation, and bootstrap tests in practice give very similar results, at least on TREC data with 50-topic sets. The randomized permutation test makes no assumptions about underlying population distributions or true deltas, and therefore does not provide the theoretical framework for power analysis. The bootstrap test takes the distribution of the sample (possibly with its mean shifted) as the best estimate of the distribution of the underlying population, and resamples with replacement from the sample to simulate sampling from the population. Bootstrapping is well suited to post-hoc calculation of power; see Efron and Tibshirani [1993, chapter 25] for more details. However, it is problematic for design-phase power estimation, as at design phase there is no sample to bootstrap from.

3.2 Power

Consider a typical scenario when an experimenter is comparing a new system a against a baseline system b. The experimenter runs the experiment, and finds $a > b$, but the result is not statistically significant. What conclusion can the experimenter draw? Post-hoc power analysis addresses the question of sensitivity. It indicates whether the test should reliably have detected a δ that the experimenter regards as consequential. If the test should have, but didn't, then the experimenter can conclude that there was no substantial effect, and try something else. (Similar conclusions can be drawn from an examination of the test's confidence interval [Hoenig and Heisey, 2001]). However, if the test neither finds significance, nor turns out to have been powerful enough to reliably detect a consequential delta, then the result is indeterminate and the value of the new method is unknown. An inconclusive experiment is all the more serious because the regimen of statistical hypothesis testing does not simply allow the experiment to be repeated with new topics, as finding significance on a second test is suspect if the test has occurred because of a failure to find significance on the first one.

In hypothesis testing, the parameter α allows the experimenter to control the risk of falsely finding a significant difference when no difference in fact exists, what is known as a *Type I* error. The converse risk, of failing to reject the null hypothesis when a difference between the systems does in fact exist, is termed a *Type II* error, and the probability of it occurring is denoted by β. To derive a value for β, one must in general posit a specific alternative hypothesis H_a, for instance, that the true δ between systems (under whatever metric is employed) is 0.07. Additionally, β depends on α (the smaller α, the greater β), the variability of the underlying population, and the size of the sample. We may also consider the probability of correctly rejecting the null hypothesis when H_a is true, namely $1 - \beta$. This value is the *power* of the statistical test. Informally, it expresses the test's ability to detect real differences. For more details, see Cohen [1988] and Hays [1991, chapter 7].

Both the significance level α to set and the power $1 - \beta$ to seek are at the experimental designer's discretion. A conventional value of α is 0.05, and a typical value of β is 0.2, giving power 0.8. These settings imply that the designer regards a false positive as being four times as serious as a false negative; the designer is willing to accept a 20% risk of missing a consequential effect if it is there, but only a 5% risk of finding a significant difference that does not exist.

In design-phase power analysis, the crucial question is, will the proposed experiment have sufficient power to reliably detect the predicted (or a consequential) effect? The main variable under the experimenter's control to try to ensure this outcome is the sample size of the proposed experiment. However, to perform the calculation, the experimenter must also have a reliable estimate of the variability of the sampled population. The greater the variability, the more likely it is to obscure any real difference between systems, regardless of the hypothesis test employed, and hence the larger the sample size that will be required to detect a given true difference.

The formula for computing statistical power depends on the statistical test that is to be invoked. The power of a one-sided paired t test, using the normal approximation, is given by:

$$P \approx \Phi\left(\sqrt{n} \cdot \frac{\delta}{\sigma} - z_{1-\alpha}\right) \tag{2}$$

where σ is the standard deviation of the underlying population, $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the normal cumulative density function or CDF (for instance, 1.644 for $\alpha = 0.05$), and Φ is the normal CDF, a monotonically increasing function. To maintain the same power (to keep the left term in the parentheses in Equation 2 constant) while halving the detectable delta, or handling twice the standard deviation, requires quadrupling the sample size.

The main practical problem facing the experimental designer when attempting to predict statistical power using a formula like Equation 2 is to estimate σ, the standard deviation of the population being sampled. In post-hoc analysis, σ is estimated directly from the observed standard deviation s of the sample, but, when designing the experiment, the sample does not yet exist. For some experiments, the standard deviation of the population is known or can be estimated. For instance, if the experiment is to measure the survival time of mice after a certain treatment, then the standard deviation of the longevity of mice is likely already known. However, the population standard deviation σ in differences in per-topic (say) AP scores between IR systems is not known. Estimating σ is the main subject of this paper.

3.3 Effect size

So far, the (minimum) effect that the experiment is attempting to detect has been defined in terms of an absolute δ, in units of whatever the evaluation metric being employed is; for instance, the researcher wishes to reliably detect a true AP improvement of 0.07

of the experimental over the baseline system. Calculating t test power in terms of an absolute effect then requires an estimate of the population's standard deviation, σ.

An alternative way to quantify the effect that the experimenter predicts or wishes to detect is in terms of *effect size* (ES) [Cohen, 1988]. There are several ways to specify ES, but the simplest is:

$$\text{ES} = \frac{\delta}{\sigma} \tag{3}$$

that is, ES is a true delta normalized by the population's standard deviation. Effect size is a unitless metric, or rather, the unit is in standard deviations, applicable to any experimental population. The concept is similar to that of standardized metrics [Webber et al., 2008]. It is even possible to generalize ES strengths. Cohen [1988], for instance, proposes that an ES of 0.8 represents a *large*, 0.5 a *medium*, and 0.2 a *small* effect. These are rough categories, and Cohen himself gives several cautionary examples. Nevertheless, they illustrate the benefits of moving to a unitless metric.

Predicting effect in terms of ES as defined in Equation 3 has a particular advantage when calculating the power of a t test. Equation 2 shows that t test power is constant (under the normal approximation) for a given ES and number of topics. Having specified topic set size, power, and α, ES is also fixed. One can say, for instance, that, using the typical 50-topic TREC test collection, a two-tailed t test with $\alpha = 0.05$ can detect an ES of 0.40 with power 0.8, regardless of the metric employed. In other words, using Cohen's categorizations, 50-topic TREC collections can reliably detect medium effects, but not small ones. Additionally, if effect can be specified in terms of ES, then it is not necessary to estimate population standard deviation in order to predict test power during experimental design, as standard deviation is already incorporated into ES.

The question is then whether it is meaningful for experimenters to specify predicted or consequential effect in terms of ES. Part of the answer lies in recognising that there are two sources of variability in topic score deltas: intrinsic variability in system performance, and extrinsic variability from the experimental setup. Intrinsic variability refers to the consistency of the experimental system's improvement over the baseline system. For the same δ, the experimental system could marginally outperform the baseline on all topics, or marginally underperform on most but significantly outperform on a few. Specifying ES instead of absolute effect can account for intrinsic variability when the experimenter cares about consistency of improvement rather than raw magnitude. Extrinsic variability refers to numerous elements of variability in the experimental setup. The experimenter will not want to simply incorporate such extrinsic variability into the specified ES, but rather wants to see through this variability to the true underlying difference.

In this paper, we assume that the researcher is quantifying effect in absolute terms, and therefore needs to estimate population standard deviation during the experimental design phase.

4. EXPERIMENTAL TOOLS AND DATA

The experimental data for this paper is drawn from the TREC test collections and the systems officially submitted to the TREC experiments. A test collection and the runs submitted to the track the collection was developed for will be termed a *test set*. The main test set used here is from the Robust Track of TREC 2004. The test collection contains 249 topics: Topics 301–450 from the AdHoc tracks of TREC-6, TREC-7, and TREC-8; Topics 601–650 from the TREC 2003 Robust Track; and Topics 651–700 newly created for the 2004 track (one of which was dropped as returning no relevant documents). Here, we only use Topics 301–450, in order to improve topic homogeneity; the Robust topics had smaller pools and so smaller estimates of the number of relevant documents R. A total of 110 systems were submitted to the Track. Of these, we exclude the 32 description-only runs. The TREC-6 topics have a peculiarity where for most topics the description is lacking one or more topic keywords from the title, leading to highly variable performance by description-only runs. Note that the methods described here were initially developed using the TREC-8 AdHoc test set, with similar experimental results.

A system's run against a topic is scored using an IR evaluation metric, a function mapping a vector of (binary or graded) relevancies to a single score. One of the most widely used metrics, and the one employed in this paper, is average precision (AP). Average precision is most easily described by building it up from its constituent parts. The *precision* of a run at depth d is the proportion of the first d documents in the run that are (binary) relevant. The sum of precisions (SP) adds up the precision at every position that a relevant document is returned in a run. Then SP is normalized to create AP by dividing by the total number of (known) relevant documents, R. The value of R is determined either by exhaustive assessment of the document corpus; or, more practically, as the number of relevant documents returned by a set of experimental systems run against the test collection when the collection was formed, as in a TREC experiment; or finally, in an experiment using new topics, as the number of relevant documents found in performing assessment for that experiment. An AP score ranges from 0 to 1, with 0 meaning no relevant documents were returned, and 1 meaning all known relevant documents were returned at the head of the run.

5. ESTIMATING DELTA DEVIATION

There are several ways in which the IR experimental designer can attempt to estimate σ, the standard deviation of score deltas for a planned experiment. One is to base the estimate on past experience. Another is to undertake a pilot experiment on a small number of topics before proceeding to the main experiment. And a third is to iteratively increase the sample size, updating the estimation at each iteration.

5.1 Based on previous experience

An experimenter might estimate expected population delta standard deviation based on previous experience. What constitutes previous experience is difficult to precisely quantify. A simple approach is to examine the standard deviations observed in past experiments for the selected metric, such as AP delta standard deviations in TREC experiments. The problem is that there is no single distribution of AP score deltas that applies to all system pairs, and the range of standard deviations across different system pairs is broad.

The 78 systems in the Robust test set form 3003 system pairs. Each system pair produces a set of AP score deltas across the 150 test topics. Figure 1 shows the distribution of standard deviations of these 3003 sets of system pair AP score deltas. The mean is 0.15, and the 95th percentile is 0.19. The first thing that Figure 1 indicates is that there is no single population of AP score deltas which all system pair deltas are drawn from. Bootstrapping indicates that a population of AP scores with a $\sigma = 0.15$ under i.i.d. sampling would give a 95% interval on sample standard deviation of around 0.12 to 0.18 for a sample size of 150 topics, whereas the 95% interval here is 0.07 to 0.20, more than twice as broad. Thus, the deltas between each pair of systems constitute a distinct population, and the σ of each such population must be separately estimated.

Past experience does not, therefore, specify a single σ of AP score deltas. Let us imagine instead that the experimental designer had available to them experience equivalent to knowing that the

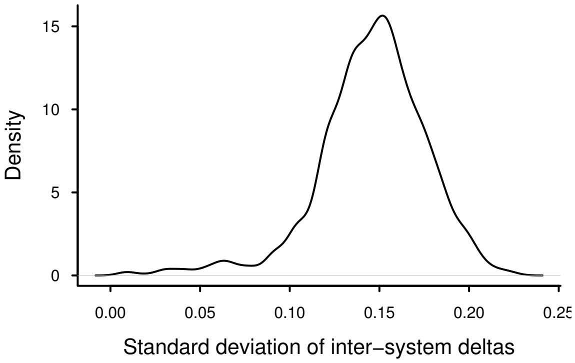

Figure 1: Density of standard deviations of between-system per-topic AP score deltas, for the TREC 2004 Robust Track runs.

two experimental systems were drawn from the Robust test set. In this case, the σ of their population of AP score deltas could be estimated at 0.15; this is only an estimate, as we know from the previous paragraph that each system pair is drawn from a distinct population and therefore has a distinct σ. The size of the true δ that the experimenter might be trying to detect will vary. A reasonable figure for trying to improve upon an established baseline is the size of the difference between the mean of the second quartile system AP scores and the mean of the first quartile, which for the test set is 0.033. To have power 0.8 given $\sigma = 0.15$ on $\delta = 0.033$ requires 164 topics. One notes immediately that the traditional 50-topic TREC collection is inadequate to reliably detect such a true difference. That is, an experiment should contain at least 150 topics if a typical top-quartile system is to be reliably distinguished from a typical second-quartile baseline.

Taking the estimate of σ as the average given by past experience means that there is a roughly 50% chance that the post-hoc analysis will show the experiment to have failed to achieve the desired power, even if past experience is a reliable guide. This is because the standard deviation of the actual sample has a 50% chance of being higher than the estimate. To be confident of achieving power, one should take a higher percentile of the empirical distribution of standard deviations. A conventional confidence level is 95%, which in our example would require taking the 95th percentile of 0.19, requiring 262 topics to reliably detect the δ in question. And, conversely, post-hoc analysis in the average case would demonstrate that this was roughly 100 topics more than was required to achieve the desired power. That is, on average, the standard deviation of the actual sample will turn out to be the empirical mean, and sufficient power would have been achieved with 164 topics.

Table 1 gives mean and 95th percentiles on AP delta standard deviations for several other TREC test sets. There is considerable variability in means and upper percentiles between test sets, with the means of some being close to the upper percentiles of others. This shows that, even in the restricted domain of TREC experiments, previous experience is an imperfect guide. In each case, the 95th percentile standard deviation is 25% to 40% more than the mean, leading under Equation 2 to 60% to 100% more topics than in the mean case, and therefore on average the same percentage more topics assessed than post-hoc analysis will show to have been necessary. Table 1 also gives the minimum AP differences detectable with power 0.8 using 50 topics with the mean and 95th percentile standard deviations. These suggest that the standard 50 topic TREC collection can only be relied on to detect true AP deltas in the range 0.06–0.08.

Test Set	AP delta σ		Detectable δ	
	Mean	95%	Mean	95%
TREC-3 AdHoc	0.144	0.198	0.058	0.080
TREC-4 AdHoc	0.171	0.220	0.069	0.089
TREC-5 AdHoc	0.170	0.241	0.069	0.097
TREC-6 AdHoc	0.196	0.259	0.079	0.105
TREC-7 AdHoc	0.152	0.207	0.061	0.084
TREC-8 AdHoc	0.160	0.226	0.065	0.091
TREC-9 Web	0.167	0.225	0.067	0.091
TREC2001 Web	0.143	0.202	0.058	0.081
TREC2004 TB	0.131	0.185	0.053	0.075
TREC2005 TB	0.142	0.191	0.057	0.077
Average	0.157	0.215	0.064	0.087

Table 1: Mean and 95th percentile of standard deviation of per-topic, between-system AP score deltas, for different TREC tracks, and delta detectable with power 0.8 using 50 topics for these standard deviations.

Of course, what has been presented here is only a rough simulation of the background knowledge that might be available to the experimental designer in attempting to estimate delta standard deviation. The designer may be aware, for instance, that the experimental system is a minor modification of the baseline one. In this case, one might anticipate lesser variability in score deltas. At the same time, however, one might be looking for a smaller absolute effect. In any case, the basic problem remains: there is no single population of AP score deltas, and therefore no single σ; past experience gives quite wide margins of error; and accounting for these margins by taking conservative estimates is expensive.

5.2 Based on trial experiments

The standard deviation of a population of score deltas can be estimated by use of trial experiments. In a trial experiment, a number of topics are sampled and assessed, and an estimation of standard deviation is made from this sample. This estimation is then used to determine the sample size for the main experiment. The trial topics should then be discarded, and an entirely new experimental set sampled anew; including the trial topics in the experimental set leads to a bias similar to that discussed ahead in Section 5.3. The experimenter might also make use of the trial experiment to determine the δ to detect, and possibly as a basis to abandon the experiment if the experimental system seems clearly worse than the baseline one.

When designing a trial experiment, one must consider, first, how many topics to include in the trial, and second, how to use the estimate of standard deviation derived. Considering the second question, the trial experiment only provides an estimate of standard deviation, and this estimate could have a wide confidence interval. As with estimates based on past experience, if the experimenter simply takes the mean as their estimate, then there is roughly a 50% chance that the true standard deviation will be higher than this. To avoid this outcome, the experimenter might choose to take an estimate from a higher percentile of the distribution; however, the higher the percentile, the more expensive the subsequent experiment will become.

The sampling distribution of the standard deviation of a normal population is itself normal, and has standard error of:

$$\sigma_S = \frac{\sigma}{\sqrt{2n}} \, . \tag{4}$$

In practice, metric deltas are not normally distributed, and AP deltas in particular are often highly skewed. Figure 2 shows the theoret-

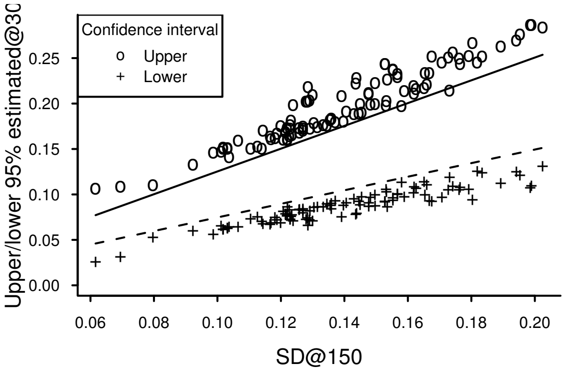

Figure 2: Theoretical (line) and empirical (point) 95% confidence interval, estimating score delta σ at 150 topics from a sample of 30, based on the TREC2004 Robust systems. The metric is AP.

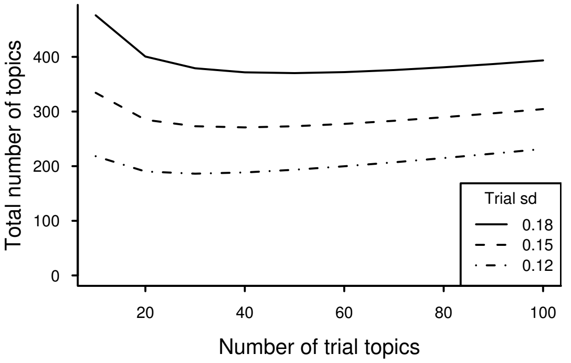

Figure 3: Total number of topics assessed following trial experiments of different sizes showing sample standard deviations of different values, where the final experiment requires power of at least 0.8 on $\delta = 0.033$ at $\alpha = 0.05$, with 95% confidence.

ical 95% confidence intervals given by Equation 4 under assumed normality, and the empirical quantiles on the TREC2004 Robust Track systems, for samples of 30 topics, indicating that the formula underestimates the variance of the estimator. Nevertheless, Equation 4 is useful to inform our discussion. Assume a trial experiment using 30 topics, producing an observed standard deviation of 0.15. The upper end of the one-tailed 95% confidence interval is then 0.183. If the experimenter required a power of 0.8 for a true delta of 0.033, then the standard deviation estimate of 0.15 would call for a topic set size of 164 topics, but there would be an even-money chance that the experiment would not achieve the desired power. On the other hand, the experimenter could take the upper end of the confidence interval at 0.183 and be confident of achieving the desired power. But this would require a topic set size of 243 topics, in addition to the 30 topics spent on the trial experiment.

Decreasing the width of the confidence interval on $\hat{\sigma}$, and therefore the size of the full experiment, can be achieved by including more topics in the trial experiment. However, from Equation 4, the expense is quadratic: halving the width of the confidence interval requires quadrupling the size of the trial, and after a while the extra size of the trial experiment is not justified by the savings in the full experiment. Figure 3 shows the full number of topics that need to be assessed in a moderately stringent experimental setup based on the trial experiment approach, for different trial experiment sizes finding different estimates of standard deviation. The optimal trial size itself depends on the standard deviation of the population – a circular problem. In any case, for each actual standard deviation, the minimum full (trial plus experimental) topic set size needed for confidence in achieving post-hoc power is 60% to 80% greater than that needed in the actual average case.

The trial method therefore has efficiency roughly equivalent to the basing of estimates on previous experience. The advantages are, first, that it requires no previous experience, and is not misled if previous experience turns out to be unreliable; and second, the trial can yield useful information, before proceeding (or not) to the full experiment.

5.3 Based on iterative estimation

In many fields of research, an experiment must be performed all at once; it is not possible, for practical or theoretical reasons, to iteratively add new subjects to the sample if the current set proves not to be large enough. However, IR research does not appear to be under such a constraint, at least when testing a limited number of implemented systems under the experimenter's control. Specifically, when calculating a test's power, if the number of topics initially chosen turns out not to provide adequate power because the standard deviation of the sample (and hence, by inference, of the population) is higher than expected, then it would seem that one could just add more topics. Such an experimental methodology is described in Algorithm 1. After each topic has been sampled and evaluated, and the two systems scored, we can check to see whether the desired power has been reached. If it has, we stop, and perform the significance test. If not, we sample another topic. In practice one would start with at least the minimum sample size reliably supported by the significance test employed.

The great advantage of the method presented in Algorithm 1 is that no assessment effort is wasted: the desired experimental power is precisely achieved with the minimum number of topics compatible with random sampling. This is in contrast to the previous methods, which attempt to estimate the standard deviation in advance, and then judge precisely that many topics. Such methods, conservatively employed, typically assess as much as twice as many topics as post-hoc prove necessary for the desired power. In addition, the iterative approach is guaranteed to achieve the desired power, whereas the estimate-then-sample approaches still allow a chance that the desired power is not achieved.

It would be a seriously flawed methodology to employ Algorithm 1 but additionally to check for significance after adding each topic, and stop when significance is found. Take the 137 Robust system pairs with an observed p value between 0.05 and 0.10 on Topics 301–450, that is, pairs that are close to but not achieving significance at level $\alpha = 0.05$. If significance is checked after every topic subset size from 50 topics up to 150, then on one random trial almost two thirds (89) of these systems pairs were found significant at level $\alpha = 0.05$ for at least one topic set size. That is, the possibility of false positives is greatly increased by such repeated testing. Thus, in any iterative approach, the stopping condition has to be specified in advance, independent of findings of significance, and strictly adhered to. It would seem, however, to be allowable

Algorithm 1 Iterative sampling

```
Input: $\delta$, the target detectable true delta
$d \leftarrow \infty, T \leftarrow \{\}$
while $d > \delta$ do
    $T \leftarrow T \cup \{sample(\mathcal{T})\}$
    $d \leftarrow$ calcDetect($T$)
end while
Perform significance test
```

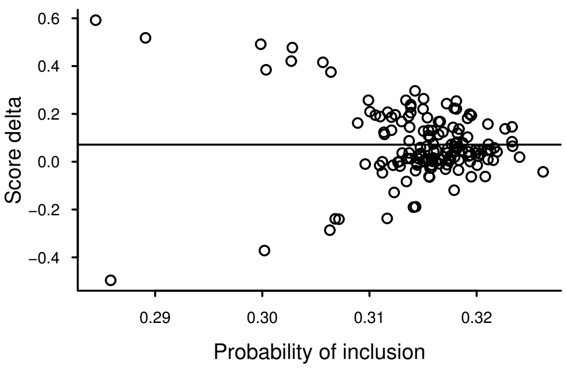

Figure 4: Probability of inclusion against per-topic delta for Topics 301–450 comparing `fub04Tge` and `polyutp1` from the TREC 2004 Robust Track, as averaged over 20,000 random trials. Each trial iteratively samples topics without replacement until a test power of 0.8 for a true delta of 0.06 is achieved.

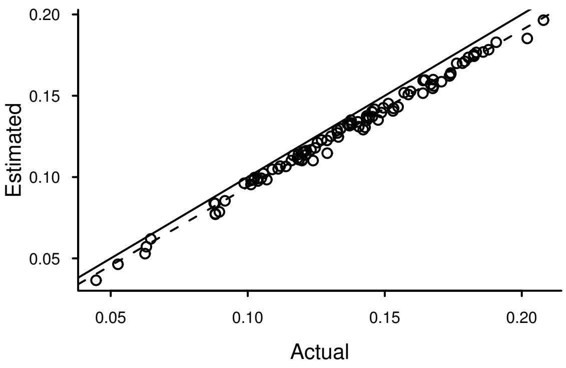

Figure 5: Mean standard deviation estimates using the iterative sampling method compared to actual standard deviation, for AP score deltas between 100 baseline and experimental system pairs drawn from the TREC2004 Robust Track systems, on Topics 301-450. The dotted line is the line of best fit, which has slope 0.965, and root mean squared residual of 0.0028.

to abandon the experiment early if it became clear that significance was not going to be achieved; that issue will not be pursued here.

Testing significance after each topic is clearly wrong; but re-estimating population standard deviation, and from that test power, either after each topic as described in Algorithm 1, or at certain intervals, might not on the surface seem to suffer from the same problems. However, iteratively re-estimating σ and hence power leads to its own, more subtle form of bias. The problem is that sampled sequences that have a lower observed standard deviation will lead to smaller subsets than those with a higher one, as they will achieve the desired power sooner. This means that the probability of inclusion (that is, the probability that a topic is in the sampled set at the time the sampling concludes with a "powerful enough" set of topics) will be higher for topics whose score delta is more typical of the population than it will be for topics with non-typical score deltas, even though the sampling probability is uniform.

Figure 4 illustrates the inclusion bias arising from iterative power calculation when seeking to compare two particular TREC 2004 Robust Track systems. Different topics have significantly different chances of being included in an iteratively sampled topic set. Those topics whose score deltas are more typical of the sample have a higher likelihood of inclusion, whereas atypical topics have a much lower one.

The effect of the inclusion bias towards topics with typical deltas is to underestimate the standard deviation of the population. The extent of this bias depends on the distribution of the underlying population, and this distribution differs for every system pair. For a given system pair, the bias can be experimentally estimated, but only post-hoc, after true random sampling. Assume that the distribution of the population $\mathcal{D}$ is identical to that observed in the sample D. That is, every $d \in \mathcal{D}$ takes on one of the values observed in D with equal probability. This is the fundamental assumption made in bootstrap statistical testing. The justification is that, although the true population distribution will not be identical to that of the sample, the sample is nevertheless (absent other information) the most likely estimate of the true distribution. Under the assumption that $\mathcal{D}$ is distributed according to D, the population standard deviation $\sigma_{\mathcal{D}}$ is known, as it is by definition the same as the sample standard deviation σ_D. Then the iterative sampling strategy can be simulated on $\mathcal{D}$ by sampling *with replacement* from D, stopping at some sample D' of m topics when the desired delta δ is detectable with power 0.8 given the observed sample standard deviation $s_{D'}$. We set δ to the value properly detectable after 100 topics, given $\sigma_{\mathcal{D}}$. The stopping $s_{D'}$ will be our estimate of $\sigma_{\mathcal{D}}$. Then, $s_{D'}$ can be compared against σ_D. Repeating this procedure will give us distributions of estimates of $\sigma_{\mathcal{D}}$ from the iterative approach, and indicate the bias of the estimator.

The intended scenario is of experimenters trying to improve on a good baseline system. For the baseline system in each experiment, therefore, we randomly select one of the systems from the second quartile of runs. The experimental system is randomly selected from the top three quartiles of runs; we assume that truly poor, fourth quartile systems have been weaned out by earlier testing.

To illustrate the method, take a random baseline-experimental system pair from the TREC 2004 Robust experimental set, with `fub04Tg` as the baseline system b and `pircRB04t1` as the experimental system a. The mean delta between a and b on Topics 301–450 is 0.036, $a > b$, and the standard deviation is 0.128. We set the delta we want to detect as the delta detectable with power 0.8 on 100 topics from the real population; this happens to be 0.036 as well. We then run 1,000 repeats of the iterative sampling approach until the desired power is reached. The iterative method takes an average of 95 topics to achieve the desired power, instead of the true 100, and the mean estimate of population standard deviation is 0.123, instead of the true value of 0.128. However, the mean of the delta means is correctly estimated at 0.036. Also, while the distribution of mean estimates is very close to normal, that of standard deviation estimates is noticeably non-normal, with a fat lower tail of low standard deviation estimates.

Figure 5 displays the mean standard deviation estimates, across 1,000 trials each, for 100 randomly selected baseline-experimental system pairs. The mean estimate of standard deviation arrived at using the iterative sampling method is consistently below the actual value, making this a biased estimator. The standard deviations are on average underestimated by 3.5% for this data set. Reducing the frequency at which power is retested only decreases the bias slightly; adding 40 topics per iteration to the topic set rather than just 1 leads to an average underestimate of 2.8%. The estimates of the mean (not shown) are unbiased.

Underestimating standard deviation but fairly estimating mean delta biases the iterative sampling method towards creating topic sets with more apparent power and a higher likelihood of finding statistical significance than would topic sets of the same size generated by pure random sampling. That this leads iterative sampling to higher rates of false positives in significance testing (finding significance where none exists) can be empirically demonstrated. As previously, take the observed distribution D of deltas for a pair of systems as the population distribution, but first shift it so that the mean of D is zero. Doing this creates the null hypothesis of paired two-tailed significance testing, that is, a population $\mathcal{D}$ of deltas with true mean zero. Any finding of significance on a sample drawn from this population is therefore by definition a false positive. Then, sample with replacement from D using the iterative sampling method until the desired power is achieved. Iterative sampling is repeated multiple times, and the proportion of iteratively sampled topic sets which find significance is recorded. Finally, true random sampling is performed, using the same topic set sizes as observed in iterative sampling, and the proportion of randomly sampled topic sets which find significance is recorded. This gives us the false positive rates for the random and iterative sampling methods on this sample of populations and topic set sizes.

Figure 6 displays the false positive rates for the iterative and random sampling methods, for 25 experimental and baseline system pairs. Here, the desired power is set, oracle-like, to that achievable after 80 topics given $\sigma_{\mathcal{D}}$, and the initial sample size for iterative sampling is 40, to satisfy the requirements of the t test on non-normal data. The mean false positive rate for true random sampling is 0.0507, roughly as expected for a significance level $\alpha = 0.05$. The mean false positive rate of the iterative method is 0.0540, some 7% higher. Additionally, the iterative method gives a higher false positive rate for 19 of the 25 populations. A two-tailed paired Wilcoxon test finds these differences significant at level 0.005.

The two rightmost points on Figure 6 merit further consideration. The higher one, where both random and iterative sampling produce high false positive rates, is for a population derived from a pair of systems from the same research group. The characteristic of this system pair is that they achieve similar scores on most topics, but one system outperforms the other markedly on a couple of topics. This reflects a common experimental situation, in which a baseline system is modified in a way that may only be evident on a few topics. Zero-centered, this creates a population of many small, mostly negative values and a couple of big, positive ones. Failing to sample the large values happens readily, whichever sampling method is used, and leads to misleading consistency and hence significance. The problem here is not with employing the t test, but with sampling itself: extreme values can be missed. Even with a sample size of 150, true random sampling still produces a 6% false positive rate. The lower of the rightmost two points in Figure 6, in contrast, is from a relatively symmetric but fat-tailed population. True random sampling is not bothered by this, but iterative sampling's bias towards including typical deltas leads to a high false-positive rate.

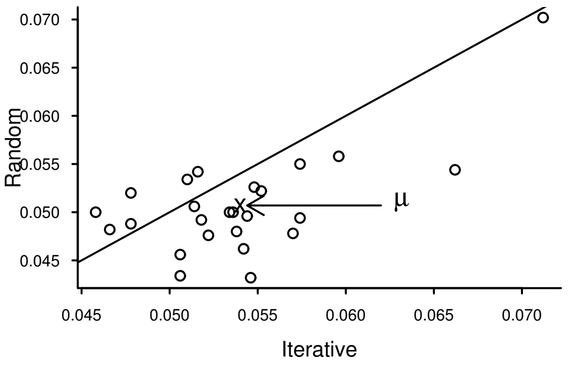

Figure 6: Proportion of false significance readings for iteratively sampled topics compared to randomly sampled topics, across 25 populations derived from randomly sampled system pairs from the TREC2004 AdHoc Track, shifted to a mean delta of 0, with $5{,}000$ samples per pair.

The iterative estimation method, therefore, while efficient in its use of topics and easy to implement, leads to a bias in favour of both experimental power and finding significance. Our experiments indicate that the bias is slight, but do not provide a general adjustment factor for it. Nevertheless, the pragmatic researcher may be prepared to use this method and note that p values produced in significance tests may be biased marginally downwards.

5.4 Suggested methodology

Faced with a pair of systems to evaluate on new topics, a researcher rich in relevance assessment resources and armed either with strong previous experience or the results of a trial experiment can proceed to make a single, conservative estimate of σ and assess the full (and generally large) set of topics necessary to be confident of achieving the desired power. A poorer but theoretically fastidious and temperamentally stoic researcher might take an average estimate and risk variability turning out to exceed that and rendering the experiment inconclusive (and, for the scrupulous, unrepeatable). Or the researcher might abandon absolute measures and express consequential effect in terms of effect size, with its attendant limitations and vagueness. Any of these approaches will enable the statistical significance of the experiment's results to be tested and reported with the minimum of caveats.

However, for a researcher with scarce assessment resources who wishes to quantify consequential effect in absolute terms and is (understandably) unwilling to risk an inconclusive experiment, we suggest a hybrid approach. The predicted or consequential δ must be stated at the outset. An initial best (non-conservative) estimate of σ should be made, either through experience and a judgment of the likely similarity of the two systems, or using a trial experiment. The indicated number of topics should then be assessed, and the systems evaluated. If desired power has not been achieved, then σ should be re-estimated as the observed sample standard deviation, and the indicated number of additional topics assessed and evaluated. (Observed standard deviation is likely to be an overestimate of population σ, since the only reason we are observing it is that it is higher, possibly by chance, than our initial estimate; however, a slightly conservative estimate here is aesthetically desirable to reduce the number of iterations.) This process is repeated until power is achieved. Then, and only then, significance can be tested for.

The proposed methodology is assessment-thrifty and guaranteed to obtain the desired power. The downside is that the reported significance is likely to be slightly exaggerated. Naturally, the researcher needs to report this fact, and also that the exact degree of bias is uncertain. The researcher also needs to state the experimental methodology employed, including the δ used to calculate power, the initial topic set size, and the number of iterations. This must be reported even if power is achieved by the initial topic set, without the need for further iterations. The only reason there weren't further iterations is because power was achieved; the subsequent significance test is not independent of this methodological choice, and will be (mildly) biased.

6. EVALUATION DEPTH

So far, evaluation effort has been calculated in terms of the number of topics that have to be assessed, if no appropriate test collection already exists. However, the true cost is in the number of documents that have to be judged, which is a function not only of the number of topics sampled, but the depth to which a metric is calculated, and documents judged, for each topic (leaving aside, for simplicity, the start-up costs per topic). By evaluating each run

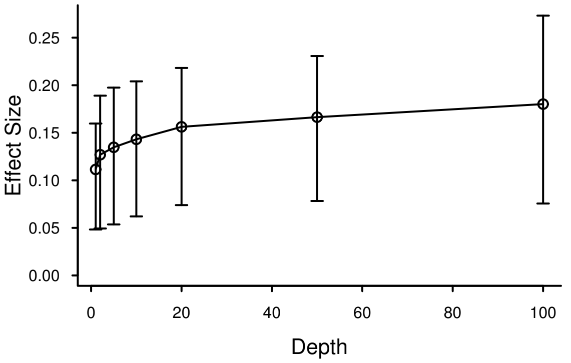

Figure 7: Observed effect size of AP at different evaluation depths. The mean effect size along with the inter-quartile range is shown. 1,000 baseline/experimental system pairs are randomly selected from the TREC2004 Robust Track systems, and observed effect size calculated at each depth.

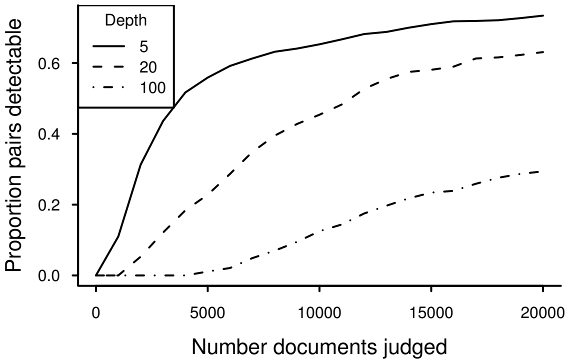

Figure 8: Proportion of empirical experimental effect sizes detectable for different number of documents judged with different assessment depths under AP. The distribution of effect sizes is taken from that observed on 1,000 baseline/experimental system pairs randomly sampled from the TREC2004 Robust Track systems. Power is 0.8, α is 0.05.

to shallower depths, less assessment effort is spent on each topic. Therefore, for the same assessment budget, it is possible to sample more topics, and a larger sample means greater statistical power. However, shallower assessment of each topic is likely to increase the variability of scores. This in turn will increase the standard deviation of score deltas, leading to a tradeoff. In this section, we address whether more statistical power results from evaluating fewer topics deeply, or more topics shallowly, given a fixed assessment budget measured in terms of documents judged. In performing this power analysis, ES will be employed rather than standard deviation directly, to account for the fact that different evaluation depths may display different delta distributions.

The number of documents that must be assessed for relevance in a paired experiment is almost linear in the depth of the evaluation. Averaging across 100 randomly selected baseline-experimental system pairs from the Robust Track experimental data set, there are 151 documents to assess for depth 100 evaluation of the two runs, 15.7 for depth 10, 8.1 for depth 5, and 3.37 for depth 2. Thus there is roughly the same document assessment effort in evaluating 50 topics to depth 100 for two runs as 900 topics to depth 5.

Figure 7 shows the range of observed ES for the pairwise experiment at different evaluation depths using the AP metric (the observed ES for a system pair is the mean divided by the standard deviation of per-topic score deltas between the two systems). For each system pair, AP scores are calculated to the specified depth, using only the known relevant documents found by the two systems to that depth. As evaluation depth is increased, score deltas do become more consistent and therefore observed ES increases, leading to both greater experimental power and likelihood of finding significance. However, the improvement is only slight.

The increase in mean ES observed in deeper evaluation is outweighed by the extra assessment effort involved; it would be far more efficient to spend the effort on more topics, as Figure 8 demonstrates. Some 5,000 documents must be judged with depth 100 assessment for any of the observed effect sizes to be reliably detected as significant, whereas after this many judgments 23% of observed effect sizes are detectable with depth 20 assessment, and 56% with depth 5 assessment. This many judgments represent 33 topics at depth 100, 161 topics at depth 20, and 617 topics at depth 5.

These results strongly suggest that shallow evaluation of many topics is preferable to deep evaluation of a few, although some caveats need to be made. First, one needs to consider how reliably the selected metric measures what the experimenter wishes to quantify, which in most cases will be user satisfaction or utility. Also, a fuller analysis of assessment effort would take into account the effort involved in topic development and an assessor's context switch between different topics (see Carterette and Smucker [2007] for one such model). Additionally, it is arguable that the shallower evaluations will have a higher proportion of extrinsic variance than the deeper ones, and that using ES to measure power masks some of this effect. Nevertheless, these caveats seem insufficient to override the conclusion that shallower and broader evaluation is more powerful for pairwise experimentation.

One reason to consider performing deeper assessments is the reusability of the test topics. If the same topics are later to be reused to test new systems on the same document corpus, then the deeper the initial assessment, the less likely it is that new systems will return unassessed documents. This, of course, is the primary reason why such deep assessments are performed on the TREC collections, and other public experimental collections, in order for them to be publicly reusable. In a private lab, however, it might be more efficient to perform shallow assessments initially, then optionally perform supplementary assessments when new documents are returned at high ranks by newly tested systems.

7. CONCLUSIONS

We have investigated the use of statistical power analysis in IR experimental design and interpretation. One of the main problems in design phase power analysis is predicting the variability of between-system score deltas. We have demonstrated that there is no single population of score deltas for any given metric, but rather a different population for each pair of systems. Estimating delta variability from past experience or from trial experiments is inexact, and establishing reasonable confidence is expensive. On the other hand, iterative re-estimation of test power leads to bias in favour of finding significance, albeit a mild one. A hybrid approach is possible, but the experimenter must be explicit about their methodology. The issue can be avoided if the experimenter is able to specify predicted or consequential effect not as an absolute delta, but normalized by standard deviation, that is, as an effect size (ES). Which option the researcher should choose depends on their particular circumstances, but we propose a hybrid approach as an efficient (if methodologically complex) default.

One of the great benefits of power analysis is that it forces the experimenter to quantify the meaning of the experiment they are planning or have carried out. Contrary to common assumption, failure to find significance does not mean that consequential differences do not exist; one must examine the power of the test (or related measures, such as the confidence interval on the result) to draw such conclusions. And before performing an experiment, the researcher should consider what size of effect they expect, and whether the proposed test will detect it, even if they do not proceed to a more formal estimation of delta standard deviation. Inconclusive experiments are the bane of the scrupulous researcher, and trying one test collection after another until some meaningful outcome is achieved is not, to say the least, methodologically sound.

For the purposes of planning experiments, having a rough estimate of a metric's typical range of delta standard deviations, and of how much a good new system might be expected to improve over a baseline, is valuable. In these terms, the 50-topic TREC collections are distinctly unpromising from a power-analysis point of view: to reliably distinguish a second quartile from a first quartile TREC system, which seems a reasonable model for experimental improvement, a set of close to 150 topics is required, at least using the AP metric. The experimenter should therefore aggregate as many such collections together as possible to boost test power, as has been done with the Robust test collection.

If the researcher chooses or is forced to develop their own topics, then power analysis strongly suggests that shallow assessment of many queries is more reliable than deep assessment of a few. However, more work needs to be done on this; the model used here to quantify assessment effort is simplistic, and other possible consequences of very shallow assessment have not been considered.

Another area for future work is evaluating different metrics in the light of power analysis. The squared impact of delta standard deviation on sample size emphasises how important it is to use a metric that minimises extrinsic delta variability; and similarly, the narrower the distribution of delta standard deviations across different system pairs, the easier the task of safely estimating experimental power in the design phase.

Acknowledgment. This work was supported by the Australian Research Council.

References

J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proc. 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 541–548, Seattle, USA, August 2006.

C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proc. 27th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 25–32, Sheffield, United Kingdom, August 2004.

B. Carterette. Robust test collections for retrieval evaluation. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 55–62, Amsterdam, the Netherlands, July 2007.

B. Carterette and M. D. Smucker. Hypothesis testing with incomplete relevance judgments. In M. J. Silvaa, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *Proc. 16th ACM Int. Conf. on Information and Knowledge Management*, pages 643–652, Lisboa, Portugal, 2007.

B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 651–658, Singapore, Singapore, July 2008.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

G. V. Cormack and T. R. Lynam. Validity and power of t-test for comparing MAP and GMAP. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 753–754, Amsterdam, the Netherlands, July 2007.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

W. L. Hays. *Statistics*. Harcourt Brace, Fort Worth, 4th edition, 1991.

J. M. Hoenig and D. M. Heisey. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55 (1):19–24, February 2001.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 2009. To appear.

A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 375–382, Amsterdam, the Netherlands, July 2007.

T. Sakai. Evaluating evaluation metrics based on the bootstrap. In S. Dumais, E. Efthimiadis, D. Hawking, and K. Järvelin, editors, *Proc. 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 525–532, Seattle, USA, August 2006.

T. Sakai. Alternatives to Bpref. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 71–78, Amsterdam, the Netherlands, July 2007.

M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 162–169, Salvador, Brazil, August 2005.

J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4):495–512, 1997.

M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In M. J. Silvaa, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *Proc. 16th ACM Int. Conf. on Information and Knowledge Management*, pages 623–632, Lisboa, Portugal, 2007.

E. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, and Sung Hyon Myaeng, editors, *Proc. 25th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 316–323, Tampere, Finland, August 2002.

W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 51–58, Singapore, Singapore, July 2008.

E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. 15th ACM Int. Conf. on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, USA, November 2006.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998.