

# Topic set size design

Tetsuya Sakai<sup>1</sup>

Received: 26 April 2015 / Accepted: 5 October 2015 / Published online: 27 October 2015  
© The Author(s) 2015. This article is published with open access at Springerlink.com

**Abstract** Traditional pooling-based information retrieval (IR) test collections typically have  $n = 50$ – $100$  topics, but it is difficult for an IR researcher to say why the topic set size should really be  $n$ . The present study provides details on principled ways to determine the number of topics for a test collection to be built, based on a specific set of statistical requirements. We employ Nagata’s three sample size design techniques, which are based on the paired  $t$  test, one-way ANOVA, and confidence intervals, respectively. These *topic set size design* methods require topic-by-run score matrices from past test collections for the purpose of estimating the within-system population variance for a particular evaluation measure. While the previous work of Sakai incorrectly used estimates of the *total* variances, here we use the correct estimates of the within-system variances, which yield slightly smaller topic set sizes than those reported previously by Sakai. Moreover, this study provides a comparison across the three methods. Our conclusions nevertheless echo those of Sakai: as different evaluation measures can have vastly different within-system variances, they require substantially different topic set sizes under the same set of statistical requirements; by analysing the tradeoff between the topic set size and the pool depth for a particular evaluation measure in advance, researchers can build statistically reliable yet highly economical test collections.

## 1 Introduction

Many modern tracks and tasks at TREC, NTCIR, CLEF and other information retrieval (IR) and information access (IA) evaluation forums inherit the basic idea of “ideal” test collections, proposed some 40 years ago by Jones and Van Rijsbergen (1975), in the form of pooling for relevance assessments. On the other hand, our modern test collections have somehow deviated substantially from their original plans in terms of the number of topics

---

✉ Tetsuya Sakai  
tetsuyasakai@acm.org

<sup>1</sup> Department of Computer Science and Engineering, Waseda University, Tokyo, Japan

we prepare (i.e., topic set size). According to Jones and Van Rijsbergen, fewer than 75 topics “are of no real value”; 250 topics “are minimally acceptable”; and more than 1000 topics “are needed for some purposes” because “real collections are large”; “statistically significant results are desirable” and “scaling up must be studied” (Jones and Van Rijsbergen 1975, p. 7). In 1979, in a report that considered the number of relevance assessments required from a statistical viewpoint, Gilbert and Jones remarked: “Since there is some doubt about the feasibility of getting 1000 requests, or the convenience of such a large set for future experiments, we consider 500 requests” (Gilbert 1979, p. C4). This is in sharp contrast to our current practice of having 50–100 topics in an IR test collection. Exceptions include the TREC Million Query track, which constructed over 1800 topics with relevance assessments by employing the *minimal test collection* and *statAP* methods (Carterette et al. 2008). However, such studies are indeed exceptions: the traditional pooling approach is still the mainstream in the IR community.

In 2009, Voorhees conducted an experiment where she randomly split 100 TREC topics in half to count discrepancies in statistically significant results, and concluded that “Fifty-topic sets are clearly too small to have confidence in a conclusion when using a measure as unstable as  $P(10)$ .”<sup>1</sup> Even for stable measures, researchers should remain skeptical of conclusions demonstrated on only a single test collection” (Voorhees 2009, p. 807). Unfortunately, there has been no clear guiding principle for determining the required number of topics for a new test collection.

The present study provides details on principled ways to determine the number of topics for a test collection to be built, based on a specific set of statistical requirements.<sup>2</sup> We employ Nagata’s three sample size design techniques, which are based on the paired  $t$  test, one-way ANOVA, and confidence intervals (CIs), respectively. These *topic set size design* methods require topic-by-run score matrices from past test collections for the purpose of estimating the within-system population variance for a particular evaluation measure. While Sakai (2014a, e) incorrectly used estimates of the *total* variances, here we use the correct estimates of the within-system variances, which yield slightly smaller topic set sizes than those reported by Sakai (2014a, e). Moreover, this study provides a comparison across the three methods. Our conclusions nevertheless echo those of Sakai (2014a, b, e): as different evaluation measures can have vastly different within-system variances, they require substantially different topic set sizes under the same set of statistical requirements; by analysing the tradeoff between the topic set size and the pool depth for a particular evaluation measure in advance, researchers can build statistically reliable yet highly economical test collections.

The remainder of this paper is organised as follows. Section 2 discusses prior art related to the present study. Section 3 describes the sample size design theory of Nagata (2003) as well as the associated Excel tools that we have made publicly available online,<sup>3</sup> and methods for estimating the within-system population variance for a particular evaluation measure. Section 4 describes six TREC test collections and runs used in our analyses, and Sect. 5 describes the evaluation measures considered. The topic-by-run matrices for all of

---

<sup>1</sup> Precision at the measurement depth of 10.

<sup>2</sup> This paper consolidates a Japanese domestic conference paper on the CI-based topic set size design method (Sakai 2014b), an ACM CIKM 2014 conference paper on the methods based on the  $t$  test and one-way ANOVA (Sakai 2014a), and an EVIA 2014 *workshop* paper that compared the three methods (Sakai 2014e). While this paper may be regarded as the full-length journal version of the EVIA workshop paper, here we correct the mistakes in the CIKM and EVIA papers, namely the use of estimates for *total* variances rather than those for *within-system* variances.

<sup>3</sup> <http://www.f.waseda.jp/tetsuya/tools.html>.

the data sets and evaluation measures used in this study are also available online<sup>4</sup>; using our Excel tools, score matrices, and the variance estimates reported in this paper, other researchers can easily reproduce our results. Section 6 reports on our topic set size design results, and Sect. 7 concludes this paper and discusses future work.

## 2 Prior art

### 2.1 Effect sizes, statistical power, and confidence intervals

In the context of comparative experiments in IR, the *p value* is the probability of observing the observed between-system difference or something even more extreme under a null hypothesis distribution. When it is smaller than a predefined *significance criterion*  $\alpha$ , then we have observed a difference that is extremely rare under the null hypothesis (i.e., the assumption that the systems are equivalent), and therefore conclude that the null hypothesis is probably incorrect. Here,  $\alpha$  is the *Type I error probability*, i.e., the probability of detecting a difference that is not real. This much is often discussed in the IR community.

Unfortunately, *effect sizes* and *statistical power* have not enjoyed the same attention in studies based on test collections, with a small number of exceptions (e.g., Carterette and Smucker 2007; Webber et al. 2008b; Nelson 1998).<sup>5</sup> A small *p* value could mean either a large *effect size* (i.e., how large the actual difference is, measured for example in standard deviation units), or a large *sample size* (i.e., we simply have a lot of topics) (Ellis 2010; Nagata 2003; Sakai 2014d). For example, suppose we have per-topic performance scores in terms of some evaluation measure *M* for systems *X* and *Y* with *n* topics  $(x_1, \dots, x_n)$  and  $(y_1, \dots, y_n)$  and hence per-topic score differences  $(d_1, \dots, d_n) = (x_1 - y_1, \dots, x_n - y_n)$ . For these score differences, the *sample mean* is given by  $\bar{d} = \sum_{j=1}^n d_j/n$  and the *sample variance* is given by<sup>6</sup>  $V = \sum_{j=1}^n (d_j - \bar{d})^2/(n - 1)$ .

Consider the test statistic  $t_0$  for a paired *t* test:

$$t_0 = \frac{\bar{d}}{\sqrt{V/n}} = \sqrt{n} \frac{\bar{d}}{\sqrt{V}}. \quad (1)$$

It is clear that if  $t_0$  is large and therefore the *p* value is small, this is either because the sample effect size  $\bar{d}/\sqrt{V}$  is large, or just because *n* is large. Hence, IR researchers should report effect sizes together with *p* values to isolate the sample size effect. The same arguments apply to other significance tests such as ANOVA (see Sect. 3.2).

The (statistical) *power* of an experiment is the probability of detecting a difference whenever there actually is one, and is denoted by  $1 - \beta$ , where  $\beta$  is the *Type II error probability*, i.e., the probability of missing a real difference. For example, when applying a two-sided paired *t* test, the probability of rejecting the null hypothesis  $H_0$  is given by<sup>7</sup>

$$Pr\{t_0 \leq -t_{inv}(\phi; \alpha)\} + Pr\{t_0 \geq t_{inv}(\phi; \alpha)\} \quad (2)$$

<sup>4</sup> <http://www.f.waseda.jp/tetsuya/CIKM2014/CIKM2014PACK.tar.gz>.

<sup>5</sup> Kelly (2009) provides an extensive discussion on effect sizes in the context of *interactive* IR studies.

<sup>6</sup> Using  $(n - 1)$  as the denominator of *V* makes this an unbiased estimate of the population variance (Nagata 2003).

<sup>7</sup> The two probabilities represent regions in the left and right tails of a *t* distribution, respectively.

where  $t_0$  is the *test statistic* computed from the observed data, and  $t_{inv}(\phi; \alpha)$  denotes the two-sided critical  $t$  value for probability  $\alpha$  with  $\phi$  degrees of freedom. Under  $H_0$  (i.e., the hypothesis that two system means are equal),  $t_0$  obeys a  $t$  distribution with  $\phi = n - 1$  degrees of freedom, where  $n$  denotes the topic set size, and Eq. (2) is exactly  $\alpha$ . Whereas, under the *alternative hypothesis*  $H_1$  (i.e., that two population system means are not equal),  $t_0$  obeys a *noncentral  $t$*  distribution (Cousineau and Laurencelle 2011; Nagata 2003), and Eq. (2) is exactly the power (i.e.,  $1 - \beta$ ). By specifying the required  $\alpha$ ,  $\beta$  and the minimum effect size for which we want to ensure the power of  $1 - \beta$ , it is possible to derive the required topic set size  $n$ . Furthermore, this approach can be extended to the case of one-way ANOVA (analysis of variance), as we shall demonstrate in Sect. 3.

Sakai (2014d) advocates the use of *confidence intervals* (CIs) along with the practice of reporting effect sizes and test statistics obtained from significance tests. CIs can be used for significance testing, and are more informative than the dichotomous reporting of whether the result is significant or not, as they provide a point estimate together with the information on how accurate that estimate might be (Cumming 2012; Ellis 2010). Soboroff (2014) compared the reliability of classical CI with three bootstrap-based CIs, and recommends the classical CI and the simple *bootstrap percentile* CI. The CI-based approach taken in this paper relies on the classical CI, which, just like the  $t$  test, assumes normal distributions.<sup>8</sup>

## 2.2 Statistical power analysis by Webber/Moffat/Zobel

Among the aforementioned studies that discussed the power of IR experiments, the work of Webber et al. (2008b) that advocated the use of statistical power in IR evaluation deserves attention here, as the present study can be regarded as an extension to their work in several aspects. Below, we highlight the contributions of, and the differences between, these studies:

- Webber et al. (2008b) were primarily concerned with building a test collection *incrementally*, by adding topics with relevance assessments one by one while checking to see if the desired power is achieved and reestimating the population variance of the performance score differences. In contrast, the present study aims to provide straight answers to questions such as: “I want to build a new test collection that guarantees certain levels of Type I and Type II error rates ( $\alpha$  and  $\beta$ ). What is the number of topics ( $n$ ) that I will have to prepare?” Researchers can simply input a set of statistical requirements to our Excel tools to obtain the answers.
- Webber et al. (2008b) considered evaluating a given *pair* of systems and thus considered the  $t$  test only. However, test collections are used to compare  $m$  ( $\geq 2$ ) systems in practice, and it is generally not correct to conduct  $t$  tests independently for every system pair, although we shall discuss exceptional situations in Sect. 3.1. If  $t$  tests are conducted multiple times, the *familywise* error rate (i.e., the probability of detecting at least one nonexistent between-system difference) amounts to  $1 - (1 - \alpha)^{m(m-1)/2}$ , assuming that all of these tests are independent of one another

<sup>8</sup> It is now known that the  $t$  test actually behaves very similarly to distribution-free, computer-based tests, namely the *bootstrap* (Sakai 2006) and *randomisation* (Smucker et al. 2007) tests, even though historically IR researchers were cautious about the use of parametric tests (Jones and Willet 1997, p. 170; Van Rijsbergen 1979, p. 247).

- (Carterette 2012; Ellis 2010).<sup>9</sup> In contrast, the present study computes the required topic set size  $n$  by considering both the  $t$  test (for  $m = 2$ ) and one-way ANOVA (for  $m \geq 2$ ), and examine the effect of  $m$  on the required  $n$  for a given set of statistical requirements. Moreover, we also consider the approach of determining  $n$  based on the width of CIs, and perform comparisons across these three methods.
- Webber et al. (2008b) examined a few methods for estimating the population variance of the performance score deltas, which include taking the 95th percentile of the observed score delta variance from past data, and conducting pilot relevance assessments. However, it is known in statistics that the population within-system variance can be estimated directly by the residual variance obtained through ANOVA, and we therefore take this more reliable approach.<sup>10</sup> Furthermore, we pool multiple variance estimates from similar data sets to enhance the reliability. As for the variance of the performance score deltas, we derive conservative estimates from our pooled variance estimates.
  - Webber et al. (2008b) considered *Average Precision* (AP) only; we examine a variety of evaluation measures for ad hoc and diversified search, with an emphasis on those that can utilise graded relevance assessments, and demonstrate that some measures require many more topics than others under the same set of statistical requirements.

### 2.3 Alternatives to classical statistics

The basis of the present study is classical significance testing (the paired  $t$  test and one-way ANOVA, to be more specific) as well as CIs. However, there are also alternative avenues for research that might help advance the state of the art in topic set size design. In particular, the *generalisability theory* (Bodoff and Li 2007; Carterette et al. 2008; Urbano et al. 2013) is somewhat akin to our study in that it also requires variance estimates from past data. Alternatives to classical significance testing include the computer-based *bootstrap* (Sakai 2006) and *randomisation tests* (Boytssov et al. 2013; Smucker et al. 2007), Bayesian approaches to hypothesis testing (Carterette 2011; Kass and Raftery 1995), and  $p_{rep}$  (probability that a replication of a study would give a result in the same direction as the original study) as an alternative to  $p$  values (Killeen 2005). These approaches are beyond the scope of the present study.

## 3 Theory

This section describes how our topic set size design methods work theoretically. Sections 3.1, 3.2 and 3.3 explain the  $t$  test based, ANOVA-based and CI-based methods, respectively.<sup>11</sup> These three methods are based on sample size design techniques of Nagata

<sup>9</sup> Ellis (2010) remarks that the Bonferroni correction to counter this familywise error rate problem “may be a bit like spending \$1000 to buy insurance for a \$500 watch.”

<sup>10</sup> Sakai (2014e) compared the 95th percentile approach with the ANOVA-based approaches. While his ANOVA-based approaches used the *total* variances instead of *residual* variances by mistake and therefore *slightly* overestimated the population within-system variances, the 95th percentile method of Webber et al. (2008b) yielded *substantially* smaller variances, which may result in topic set sizes that are too optimistic. As Ellis (2010) recommends, we prefer to “err on the side of oversampling.”

<sup>11</sup> For completeness, Sects. 3.1 and 3.2 repeat the descriptions of the  $t$  test-based and ANOVA-based methods originally presented in Sakai (2014a); Sect. 3.3 repeats the description of the CI-based method originally presented in Sakai (2014b).

(2003). As these methods require estimates of within-system variances, Sect. 3.4 describes how we obtain them from past data. If the reader is not familiar with statistical power and effect sizes, a good starting point would be the book by Ellis (2010); also, the book by Kelly (2009) discusses these topics as well as ANOVA in the context of interactive IR.

### 3.1 Topic set size design based on the paired $t$ test

As was mentioned in Sect. 2.2, if the researcher is interested in the differences between every system pair, then conducting  $t$  tests multiple times is not the correct approach; an appropriate *multiple comparison procedure* (Boytsov et al. 2013; Carterette 2012; Nagata 1998) should be applied in order to avoid the aforementioned familywise error rate problem. However, there are also cases where applying the  $t$  test multiple times is the correct approach to take even when there are more than two systems ( $m > 2$ ) (Nagata 1998). For example, if the objective of the experiment is to show that a new system  $Z$  is better than *both* baselines  $X$  and  $Y$  (rather than to show that  $Z$  is *either* better than  $X$  or better than  $Y$ ), then what we want to ensure is that the probability of incorrectly rejecting both of the null hypotheses is no more than  $\alpha$  (rather than that of incorrectly rejecting at least one of them). In this case, it is correct to apply a  $t$  test for systems  $Z$  and  $X$ , and one for systems  $Z$  and  $Y$ .

Let  $t$  be a random variable that obeys a  $t$  distribution with  $\phi$  degrees of freedom; let  $t_{inv}(\phi; \alpha)$  denote the two-sided critical  $t$  value for significance criterion  $\alpha$  (i.e.,  $Pr\{|t| \geq t_{inv}(\phi; \alpha)\} = \alpha$ ).<sup>12</sup> Under  $H_0$ , the test statistic  $t_0$  (Eq. 1 in Sect. 2) obeys a  $t$  distribution with  $\phi = n - 1$  degrees of freedom. Given  $\alpha$ , we reject  $H_0$  if  $|t_0| \geq t_{inv}(\phi; \alpha)$ , because that means we have observed something extremely unlikely if  $H_0$  is true. (The  $p$  value is the probability of observing  $t_0$  or something more extreme under  $H_0$ .) Thus, the probability of Type I error (i.e., “finding” a difference that does not exist) is exactly  $\alpha$  by construction. Whereas, the probability of Type II error (i.e., missing a difference that actually exists) is denoted by  $\beta$ , and therefore the *statistical power* (i.e., the ability to detect a real difference) is given by  $1 - \beta$ . Put another way,  $\alpha$  is the probability of rejecting  $H_0$  when  $H_0$  is true, while the power is the probability of rejecting  $H_0$  when  $H_1$  is true. In either case, the probability of rejecting  $H_0$  is given by

$$\begin{aligned} &Pr\{t_0 \leq -t_{inv}(\phi; \alpha)\} + Pr\{t_0 \geq t_{inv}(\phi; \alpha)\} \\ &= Pr\{t_0 \leq -t_{inv}(\phi; \alpha)\} + 1 - Pr\{t_0 \leq t_{inv}(\phi; \alpha)\}. \end{aligned} \tag{3}$$

Under  $H_0$ , Eq. (3) amounts to  $\alpha$ , where  $t_0$  (Eq. 1) obeys a (central)  $t$  distribution as mentioned above. Under  $H_1$ , Eq. (3) represents the power ( $1 - \beta$ ), where  $t_0$  obeys a *noncentral*  $t$  distribution with  $\phi = n - 1$  degrees of freedom and a *noncentrality parameter*  $\lambda_t = \sqrt{n}\Delta_t$ . Here,  $\Delta_t$  is a simple form of *effect size*, given by:

$$\Delta_t = \frac{\mu_X - \mu_Y}{\sqrt{\sigma_t^2}} = \frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}} \tag{4}$$

where  $\sigma_t^2 = \sigma_X^2 + \sigma_Y^2$  is the population variance of the score differences. Thus,  $\Delta_t$  quantifies the difference between  $X$  and  $Y$  in *standard deviation units* of any given evaluation measure.

<sup>12</sup> T.INV.2T( $\alpha, \phi$ ) with Microsoft Excel 2013.

While computations involving a noncentral  $t$  distribution can be complex, a normal approximation is available: let  $t'$  denote a random variable that obeys the aforementioned noncentral  $t$  distribution; let  $u$  denote a random variable that obeys  $N(0, 1^2)$ . Then<sup>13</sup>:

$$Pr\{t' \leq w\} \approx Pr\left\{u \leq \frac{w(1 - 1/4\phi) - \lambda_t}{\sqrt{1 + w^2/2\phi}}\right\}. \tag{5}$$

Hence, given the topic set size  $n$ , the effect size  $\Delta_t$  and the significance criterion  $\alpha$ , the power can be computed from Eqs. (3) and (5) as (Nagata 2003):

$$1 - \beta \approx Pr\left\{u \leq \frac{(-w)(1 - 1/4(n - 1)) - \sqrt{n}\Delta_t}{\sqrt{1 + (-w)^2/2(n - 1)}}\right\} + 1 - Pr\left\{u \leq \frac{w(1 - 1/4(n - 1)) - \sqrt{n}\Delta_t}{\sqrt{1 + w^2/2(n - 1)}}\right\} \tag{6}$$

where  $w = t_{inv}(n - 1; \alpha)$ . But what we are more interested in is: given  $(\alpha, \beta, \Delta_t)$ , what is the required  $n$ ?

Under  $H_0$ , we know that  $\Delta_t = 0$  (see Eq. 4). However, under  $H_1$ , all we know is that  $\Delta_t \neq 0$ . In order to require that an experiment has a statistical power of  $1 - \beta$ , a *minimum detectable effect*  $min\Delta_t$  must be specified in advance: we correctly reject  $H_0$  with  $100(1 - \beta)$  % confidence whenever  $|\Delta_t| \geq min\Delta_t$ . That is, we should not miss a real difference if its effect size is  $min\Delta_t$  or larger. Cohen calls  $min\Delta_t = 0.2$  a *small* effect,  $min\Delta_t = 0.5$  a *medium* effect, and  $min\Delta_t = 0.8$  a *large* effect (Cohen 1988; Ellis 2010).<sup>14</sup>

Let  $z_P$  denote the one-sided critical  $z$  value of  $u$  ( $\sim N(0, 1^2)$ ) for probability  $P$  (i.e.,  $Pr\{u \geq z_P\} = P$ ). Given  $(\alpha, \beta, min\Delta_t)$ , it is known that the required topic set size  $n$  can be approximated by (Nagata 2003):

$$n \approx \left(\frac{z_{\alpha/2} - z_{1-\beta}}{min\Delta_t}\right)^2 + \frac{z_{\alpha/2}^2}{2}. \tag{7}$$

For example, if we let  $(\alpha, \beta, min\Delta_t) = (.05, .20, .50)$  [i.e., *Cohen’s five-eighty convention* (Cohen 1988; Ellis 2010) with Cohen’s *medium* effect],

$$n \approx \left(\frac{1.960 - (-.842)}{.50}\right)^2 + \frac{1.960^2}{2} = 33.3. \tag{8}$$

As this is only an approximation, we need to check that the desired power is actually achieved with an integer  $n$  close to 33.3. Suppose we let  $n = 33$ . Then, by substituting  $w = t_{inv}(33 - 1; .05) = 2.037$  and  $\Delta_t = min\Delta_t = .50$  to Eq. (6), we obtain:

$$1 - \beta \approx Pr\{u \leq -4.742\} + 1 - Pr\{u \leq -.825\} = .795 \tag{9}$$

which means that the desired power of 0.8 is not quite achieved. So we let  $n = 34$ , and the achieved power can be computed similarly:  $1 - \beta = .808$ . Therefore  $n = 34$  is the topic set size we want.

<sup>13</sup> Appendix 1 briefly describes how Nagata (2003) obtained this approximation.

<sup>14</sup> It should be noted that the effect sizes for paired tests (i.e., the ones discussed in the present study) and those for unpaired (i.e., two-sample) tests are not directly comparable (Okubo and Okada 2012).

**Table 1** Topic set sizes for  $(\alpha, \beta, \min\Delta_t)$

$\alpha$	$\min\Delta_t$	$\beta = .10$	$\beta = .20$
.01	.1	1492	1172
	.2	376	296
	.5	63	51
	1.0	19	16
.05	.1	1053	787
	.2	265	199
	.5	44	34
	1.0	13	107

Our Excel tool `samplesizeTTEST` automates the above procedure for any given combination of  $(\alpha, \beta, \min\Delta_t)$ . Table 1 shows the required topic set sizes for the paired  $t$  test for some typical combinations. For example, under Cohen’s five-eighty convention  $(\alpha = .05, \beta = .20)$ ,<sup>15</sup> if we want the minimum detectable effect to be  $\min\Delta_t = .2$  (i.e., one-fifth of the score-difference standard deviation), we need  $n = 199$  topics.

The above approach starts by requiring a  $\min\Delta_t$ , which is independent of the evaluation method (i.e., the measure, pool depth and the measurement depth). However, researchers may want to require a minimum detectable *absolute difference*  $\min D_t$  in terms of a particular evaluation measure instead (e.g., “I want high power guaranteed whenever the true absolute difference in mean AP is 0.05 or larger.”). In this case, instead of setting a minimum ( $\min\Delta_t$ ) for Eq. (4), we can set a minimum ( $\min D_t$ ) for the *numerator* of Eq. (4): we guarantee a power of  $1 - \beta$  whenever  $|\mu_X - \mu_Y| \geq \min D_t$ . To do this, we need an estimate  $\hat{\sigma}_t^2$  of the variance  $\sigma_t^2 (= \sigma_X^2 + \sigma_Y^2)$ , so that we can convert  $\min D_t$  to  $\min\Delta_t$  simply as follows:

$$\min\Delta_t = \frac{\min D_t}{\sqrt{\hat{\sigma}_t^2}} \tag{10}$$

After this conversion, the aforementioned procedure starting with Eq. (7) can be applied. Our tool `samplesizeTTEST` has a separate sheet for computing  $n$  from  $(\alpha, \beta, \min D_t, \hat{\sigma}_t^2)$ ; how to obtain  $\hat{\sigma}_t^2$  from past data is discussed in Sect. 3.4.

### 3.2 Topic set size design based on one-way ANOVA

This section discusses how to set the topic set size  $n$  when we assume that there are  $m \geq 2$  systems to be compared using one-way ANOVA. Let  $x_{ij}$  denote the score of the  $i$ -th system for topic  $j$  in terms of some evaluation measure; we assume that  $\{x_{ij}\}$  are independent and that  $x_{ij} \sim N(\mu_i, \sigma^2)$ . That is,  $x_{ij}$  obeys a normal distribution with a population system mean  $\mu_i$  and a common system variance  $\sigma^2$ . The assumption that  $\sigma^2$  is common across systems is known as the the *homoscedasticity* assumption<sup>16</sup>; note that we did not rely on this assumption when we discussed the paired  $t$  test. We define the population grand mean  $\mu$  and the  $i$ -th *system effect*  $a_i$  (i.e., how the  $i$ -th system differs from  $\mu$ ) as follows:

<sup>15</sup> Note that this convention, which implies that a Type I error is four times as serious as a Type II error, is only a convention (Ellis 2010). Researchers should consider whether this is appropriate for their experiments, and should not follow it blindly.

<sup>16</sup> Carterette (2012) demonstrates that the homoscedasticity assumption does not actually hold in the context of IR evaluation. However, the present study assumes that ANOVA is of some use to IR evaluation, as it is a fact that it is used by IR researchers (though not as often as the  $t$  test).



$$\mu = \frac{1}{m} \sum_{i=1}^m \mu_i, \quad a_i = \mu_i - \mu \tag{11}$$

where  $\sum_{i=1}^m a_i = \sum_{i=1}^m (\mu_i - \mu) = \sum_{i=1}^m \mu_i - m\mu = 0$ . The null hypothesis for the ANOVA is  $H_0 : \mu_1 = \dots = \mu_m$  (or  $a_1 = \dots = a_m = 0$ ) while the alternative hypothesis  $H_1$  is that at least one of the system effects is not zero. That is, while the null hypothesis of the  $t$  test is that two systems are equally effective in terms of the population means, that of ANOVA is that *all* systems are equally effective.

The basic statistics that we compute for the ANOVA are as follows. The sample mean for system  $i$  and the sample grand mean are given by:

$$\bar{x}_{i\bullet} = \frac{1}{n} \sum_{j=1}^n x_{ij}, \quad \bar{x} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n x_{ij}. \tag{12}$$

The total variation, which quantifies how each  $x_{ij}$  differs from the sample grand mean, is given by:

$$S_T = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2. \tag{13}$$

It is easy to show that  $S_T$  can be decomposed into *between-system* and *within-system variations*  $S_A$  and  $S_E$  (i.e.,  $S_T = S_A + S_E$ ), where

$$S_A = n \sum_{i=1}^m (\bar{x}_{i\bullet} - \bar{x})^2, \quad S_E = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i\bullet})^2. \tag{14}$$

The corresponding degrees of freedom are  $\phi_A = m - 1$ ,  $\phi_E = m(n - 1)$ . Also, let  $V_A = S_A/\phi_A$ ,  $V_E = S_E/\phi_E$  for later use.

Let  $F$  be a random variable that obeys an  $F$  distribution with  $(\phi_A, \phi_E)$  degrees of freedom; let  $F_{inv}(\phi_A, \phi_E; \alpha)$  denote the critical  $F$  value for probability  $\alpha$  (i.e.,  $Pr\{F \geq F_{inv}(\phi_A, \phi_E; \alpha)\} = \alpha$ ).<sup>17</sup> Under  $H_0$ , the test statistic  $F_0$  defined below obeys a (central)  $F$  distribution with  $(\phi_A, \phi_E)$  degrees of freedom:

$$F_0 = \frac{V_A}{V_E} = \frac{m(n - 1)S_A}{(m - 1)S_E}. \tag{15}$$

Given a significance criterion  $\alpha$ , we reject  $H_0$  if  $F_0 \geq F_{inv}(\phi_A, \phi_E; \alpha)$ . From Eq. (15), it can be observed that  $H_0$  is rejected if the between-system variation  $S_A$  is large compared to the within-system variation  $S_E$ , or simply if the sample size  $n$  is large. Again, the  $p$  value does not tell us which is the case.

The probability of rejecting  $H_0$  is given by

$$Pr\{F_0 \geq F_{inv}(\phi_A, \phi_E; \alpha)\} = 1 - Pr\{F_0 \leq F_{inv}(\phi_A, \phi_E; \alpha)\}. \tag{16}$$

Under  $H_0$ , Eq. (16) amounts to  $\alpha$  by construction, where  $F_0$  obeys a (central)  $F$  distribution as mentioned above. Under  $H_1$ , Eq. (16) represents the power  $(1 - \beta)$ , where  $F_0$  obeys a *noncentral*  $F$  distribution (Nagata 2003; Patnaik 1949) with  $(\phi_A, \phi_B)$  degrees of freedom and a *noncentrality parameter*  $\lambda$ , such that

<sup>17</sup> F.INV.RT( $\alpha, \phi_A, \phi_E$ ) with Microsoft Excel 2013.

$$\lambda = n\Delta, \quad \Delta = \frac{\sum_{i=1}^m a_i^2}{\sigma^2} = \frac{\sum_{i=1}^m (\mu_i - \mu)^2}{\sigma^2}. \tag{17}$$

Thus  $\Delta$  measures the total system effects in *variance units*.

While computations involving a noncentral  $F$  distribution can be complex, a normal approximation is available: let  $F'$  denote a random variable that obeys the aforementioned noncentral  $F$  distribution; let  $u \sim N(0, 1^2)$ . Then<sup>18</sup>:

$$Pr\{F' \leq w\} \approx Pr\left\{u \leq \frac{\sqrt{\frac{w}{\phi_E}}\sqrt{2\phi_E - 1} - \sqrt{\frac{c_A}{\phi_A}}\sqrt{2\phi_A^* - 1}}{\sqrt{\frac{c_A}{\phi_A} + \frac{w}{\phi_E}}}\right\} \tag{18}$$

where

$$c_A = \frac{m - 1 + 2n\Delta}{m - 1 + n\Delta}, \quad \phi_A^* = \frac{(m - 1 + n\Delta)^2}{m - 1 + 2n\Delta}. \tag{19}$$

Hence, given  $(n, \Delta, \alpha)$ , the power  $(1 - \beta)$  can be computed from Eqs. (16)–(19) as (Nagata 2003):

$$1 - Pr\left\{u \leq \frac{\sqrt{\frac{w}{m(n-1)}}\sqrt{2m(n-1) - 1} - \sqrt{\frac{c_A}{m-1}}\sqrt{2\phi_A^* - 1}}{\sqrt{\frac{c_A}{m-1} + \frac{w}{m(n-1)}}}\right\} \tag{20}$$

where  $w = F_{inv}(m - 1, m(n - 1); \alpha)$ . But what we are more interested in is: given  $(\alpha, \beta, \Delta)$ , what is the required  $n$ ?

Under  $H_0$ , we know that  $\Delta = 0$  (see Eq. 17). However, under  $H_1$ , all we know is that  $\Delta \neq 0$ . In order to require that an experiment has a statistical power of  $1 - \beta$ , a minimum detectable delta  $min\Delta$  must be specified in advance. Let us require that we correctly reject  $H_0$  with  $100(1 - \beta)\%$  confidence whenever the *range* of the population means ( $D = \max_i a_i - \min_i a_i$ ) is at least as large as a specified value ( $min D$ ). That is, we want to detect a true difference whenever the difference between the population mean of the *best* system and that of the *worst* system is at least  $minD$ . Now, let us define  $min\Delta$  as follows:

$$min\Delta = \frac{minD^2}{2\sigma^2}. \tag{21}$$

Then, since  $\sum_{i=1}^m a_i^2 \geq \frac{D^2}{2}$  holds,<sup>19</sup> it follows that

$$\Delta = \frac{\sum_{i=1}^m a_i^2}{\sigma^2} \geq \frac{D^2}{2\sigma^2} \geq \frac{minD^2}{2\sigma^2} = min\Delta. \tag{22}$$

That is,  $\Delta$  is bounded below by  $min\Delta$ . Hence, although specifying  $min D$  does not uniquely determine  $\Delta$  (as  $\Delta$  depends on systems other than the best and the worst ones), we can plug in  $\Delta = min\Delta$  to Eqs. (19) and (20) to obtain the worst-case estimate of the power.

Unfortunately, no closed formula similar to Eq. (7) is available for ANOVA. However, from Eqs. (17) and (21), note that the worse-case estimate of  $n$  can be obtained as follows:

<sup>18</sup> Appendix 2 briefly describes how Nagata (2003) obtained this approximation.

<sup>19</sup> Let  $A = \max_i a_i$  and  $a = \min_i a_i$ . Then  $D^2/2 = (A^2 + a^2 - 2Aa)/2 \leq A^2 + a^2 \leq \sum_{i=1}^m a_i^2$ . The equality holds when  $A = D/2, a = -D/2$  and  $a_i = 0$  for all other systems.

**Table 2** Linear approximation of  $\lambda$ , the noncentrality parameter of a noncentral  $\chi^2$  distribution (Nagata 2003)

$\alpha$	$\beta$	Formula
.01	.10	$\lambda = 10.439 + 5.213\sqrt{\phi_A}$
.01	.20	$\lambda = 7.736 + 4.551\sqrt{\phi_A}$
.05	.10	$\lambda = 7.049 + 4.244\sqrt{\phi_A}$
.05	.20	$\lambda = 4.860 + 3.584\sqrt{\phi_A}$

$$n = \frac{\lambda}{\min\Delta} = \frac{2\sigma^2\lambda}{\min D^2}. \tag{23}$$

To use Eq. (23), we need the  $\lambda$ . (How to obtain  $\hat{\sigma}^2$ , the estimate of  $\sigma^2$ , is discussed in Sect. 3.4.) Recall that, under  $H_1$ , Eq. (16) represents the power  $(1 - \beta)$  where  $F_0$  obeys a noncentral  $F$  distribution with  $(\phi_A, \phi_E)$  degrees of freedom and the noncentrality parameter  $\lambda$ . By letting  $\phi_E = m(n - 1) \approx \infty$ , the power can be approximated by:

$$Pr\{F_0 \geq F_{inv}(\phi_A, \infty; \alpha)\} = Pr\{\chi^2 \geq \chi_{inv}^2(\phi_A; \alpha)\} \tag{24}$$

where  $\chi^2$  is a random variable that obeys a noncentral  $\chi^2$  distribution with  $\phi_A$  degrees of freedom whose noncentrality parameter is  $\lambda$ , and  $\chi_{inv}^2(\phi; P)$  is the critical  $\chi^2$  value for probability  $P$  of a random variable that obeys a (central)  $\chi^2$  distribution with  $\phi$  degrees of freedom (i.e.,  $Pr\{\chi^2 \geq \chi_{inv}^2(\phi; P)\} = P$ ). For noncentral  $\chi^2$  distributions, some linear approximations of  $\lambda$  are available, as shown in Table 2 (Nagata 2003). Hence an initial estimate of  $n$  given  $(\alpha, \beta, \min D, \hat{\sigma}^2, m)$  can be obtained as shown below.

Suppose we let  $(\alpha, \beta, \min D, m) = (.05, .20, .5, 3)$  and that we obtained  $\hat{\sigma}^2 = .5^2$  from past data so that  $\min\Delta = \frac{\min D^2}{2\sigma^2} = .5^2 / (2 * .5^2) = .5$ . Then  $\phi_A = m - 1 = 2$  and  $\lambda = 4.860 + 3.584 * \sqrt{2} = 9.929$  and hence  $n = \lambda / \min\Delta = 19.9$ . If we let  $n = 19$ , then  $\phi_E = 3(19 - 1) = 54$ ,  $w = F_{inv}(2, 54; .05) = 3.168$ . From Eq. (19),  $c_A = 1.826$ ,  $\phi_A^* = 6.298$ , and from Eq. (20), the achieved power is  $1 - Pr\{u \leq -.809\} = .791$ , which does not quite satisfy the desired power of 80 %. On the other hand, if  $n = 20$ , the achieved power can be computed similarly as .813. Hence  $n = 20$  is what we want. Our Excel tool `sampleSizeANOVA` automates the above procedure for given  $(\alpha, \beta, \min D, \hat{\sigma}^2, m)$ .<sup>20</sup>

Recall that the  $H_1$  for ANOVA says: “there is a difference somewhere among the  $m$  systems,” which may not be very useful in the context of test-collection-based studies: we usually want to know exactly where the differences are. If the researcher is interested in obtaining a  $p$  value for every system pair, then she should conduct a multiple comparison procedure from the outset. Contrary to popular beliefs, it is generally *incorrect* to first conduct ANOVA and then conduct a multiple comparison test only if the null hypothesis for the ANOVA is rejected. This practice of sequentially conducting different tests suffers from a problem similar to that of the aforementioned familywise error rate (Nagata 1998).<sup>21</sup> An example of a proper multiple comparison procedure would be Tukey’s HSD (Honestly Significant Differences) test, its randomised version (Carterette 2012; Sakai 2014d), or the Holm–Bonferroni adjustment of  $p$  values (Boytsov et al. 2013); such a test

<sup>20</sup> While `sampleSizeTTEST` handles arbitrary values of  $(\alpha, \beta)$ , `sampleSizeANOVA` can only handle the four combinations shown in Table 2.

<sup>21</sup> Exceptions are when the ANOVA is part of the multiple comparison procedure (Nagata 1998).

should be applied directly without conducting ANOVA at all. Ideally, we would like to discuss topic set size design based on a multiple comparison procedure, but this is an open problem even in statistics. In fact, the very notion of power has several different interpretations in the context of multiple comparison procedures (Nagata 1998). Nevertheless, since some researchers do use ANOVA for comparing  $m$  systems, how the required topic set size  $n$  grows with  $m$  probably deserves some attention.

### 3.3 Topic set size design based on CIs

To build a CI for the difference between systems  $X$  and  $Y$ , we model the performance scores (assumed independent) as follows:

$$x_i = \mu_X + \gamma_i + \varepsilon_{Xi}, \quad \varepsilon_{Xi} \sim N(0, \sigma_X^2), \tag{25}$$

$$y_i = \mu_Y + \gamma_i + \varepsilon_{Yi}, \quad \varepsilon_{Yi} \sim N(0, \sigma_Y^2) \tag{26}$$

where  $\gamma_i$  represents the topic effect and  $\mu_\bullet, \sigma_\bullet^2$  represent the population mean and variance for  $X, Y$ , respectively ( $i = 1, \dots, n$ ). This is in fact just an alternative way of presenting the assumptions behind the paired  $t$  test (Sect. 3.1). To cancel out  $\gamma_i$ , let

$$d_i = x_i - y_i = \mu_X - \mu_Y + \varepsilon_{Xi} - \varepsilon_{Yi} \tag{27}$$

so that  $d_i \sim N(\mu, \sigma_d^2), \mu = \mu_X - \mu_Y, \sigma_d^2 = \sigma_X^2 + \sigma_Y^2$ . It then follows that  $t = \frac{\bar{d} - \mu}{\sqrt{V/n}}$  obeys a  $t$  distribution with  $\phi = n - 1$  degrees of freedom, where  $\bar{d} = \sum_{i=1}^n d_i/n$  and  $V = \sum_{i=1}^n (d_i - \bar{d})^2/(n - 1)$  as before. Hence, for a given significance criterion  $\alpha$ , the following holds:

$$Pr\{-t_{inv}(\phi; \alpha) \leq t \leq t_{inv}(\phi; \alpha)\} = 1 - \alpha. \tag{28}$$

Hence,

$$Pr\{\bar{d} - MOE \leq \mu \leq \bar{d} + MOE\} = 1 - \alpha \tag{29}$$

where the *margin of error* (MOE) is given by:

$$MOE = t_{inv}(\phi; \alpha)\sqrt{V/n}. \tag{30}$$

Thus, Eq. (29) shows that the  $100(1 - \alpha) \%$  CI for the difference in population means ( $\mu = \mu_X - \mu_Y$ ) is given by  $[\bar{d} - MOE, \bar{d} + MOE]$ . This much is very well known.

Let us consider the approach of determining the topic set size  $n$  by *requiring* that  $2MOE \leq \delta$ : that is, the CI of the difference between  $X$  and  $Y$  should be no larger than some constant  $\delta$ . This ensures that experiments using the test collection will be conclusive wherever possible: for example, note that a wide CI that includes zero implies that we are very unsure as to whether systems  $X$  and  $Y$  actually differ. Since  $MOE$  (Eq. 30) contains a random variable  $V$ , we actually impose the above requirement on the *expectation* of  $2MOE$ :

$$E(2MOE) = 2t_{inv}(\phi; \alpha)\frac{E(\sqrt{V})}{\sqrt{n}} \leq \delta. \tag{31}$$

Now, it is known that<sup>22</sup>

$$E(\sqrt{V}) = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}\sigma_t \tag{32}$$

where  $\sigma_t = \sqrt{\sigma_x^2 + \sigma_y^2}$  and  $\Gamma(\bullet)$  is the gamma function.<sup>23</sup> By substituting Eq. (32) to Eq. (31), the requirement can be rewritten as:

$$\frac{t_{inv}(n-1; \alpha)\Gamma(\frac{n}{2})}{\sqrt{n(n-1)}\Gamma(\frac{n-1}{2})} \leq \frac{\delta}{2\sqrt{2}\sigma_t} \tag{33}$$

In order to find the smallest  $n$  that satisfies Eq. (33), we first consider an “easy” case where the population variance  $\sigma_t^2$  is known. In this case, the MOE is given by (cf. Eq. 30):

$$MOE_z = z_{\alpha/2}\sqrt{\sigma_t^2/n} \tag{34}$$

where  $z_p$  denotes the one-sided critical  $z$  value for probability  $P$ .<sup>24</sup> By requiring that  $2MOE_z \leq \delta$ , we can obtain a tentative topic set size  $n'$ :

$$n' \geq \frac{4z_{\alpha/2}^2\sigma_t^2}{\delta^2} \tag{35}$$

First, the smallest integer that satisfies Eq. (35) can be tested to see if it also satisfies Eq. (33);  $n'$  is incremented until it does. The resultant  $n = n'$  is the topic set size we want.

Our Excel tool `samplesizeCI` automates the above procedure to find the required sample size  $n$ , for any given combination of  $(\alpha, \delta, \hat{\sigma}_t^2)$ . How to obtain the variance estimate  $\hat{\sigma}_t^2$  from past data is discussed below.

### 3.4 Estimating population within-system variances

As was explained above, our  $t$ -based and CI-based topic set size design methods require an estimate of the population variance of the difference between two systems  $\sigma_t^2 = \sigma_x^2 + \sigma_y^2$ , and our ANOVA-based method requires an estimate of the population within-system variance  $\sigma^2$  under the homoscedasticity assumption.

Let  $C$  be an existing test collection and  $n_C$  be the number of topics in  $C$ ; let  $m_C$  be the number of runs whose performances with  $C$  in terms of some evaluation measure are known, so that we have an  $n_C \times m_C$  topic-by-run matrix  $\{x_{ij}\}$  for that evaluation measure. There are two simple ways to estimate  $\hat{\sigma}^2$  from such data. One is to use the residual variance from one-way ANOVA (see Sect. 3.2):

$$\hat{\sigma}_C^2 = V_E = \frac{\sum_{i=1}^{m_C} \sum_{j=1}^{n_C} (x_{ij} - \bar{x}_{i\bullet})^2}{m_C(n_C - 1)} \tag{36}$$

<sup>22</sup> Note that  $\sqrt{V}$  is not an unbiased estimate of  $\sigma_t$  while  $V$  is an unbiased estimate of  $\sigma_t^2$  (i.e.,  $E(V) = \sigma_t^2$ ) (Nagata 2003; Okubo and Okada 2012).

<sup>23</sup> GAMMA( $\bullet$ ) with Microsoft Excel 2013.

<sup>24</sup> NORM.S.INV( $1 - P$ ) with Microsoft Excel 2013.

where  $\bar{x}_{i\bullet} = \frac{1}{n_C} \sum_{j=1}^{n_C} x_{ij}$  (sample mean for system  $i$ ).<sup>25</sup> The other is to use the residual variance from *two-way ANOVA without replication*, which utilises the fact that the scores  $x_{\bullet j}$  for topic  $j$  correspond to one another:

$$\hat{\sigma}_C^2 = V_E = \frac{\sum_{i=1}^{m_C} \sum_{j=1}^{n_C} (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2}{(m_C - 1)(n_C - 1)} \tag{37}$$

where  $\bar{x}_{\bullet j} = \frac{1}{m_C} \sum_{i=1}^{m_C} x_{ij}$  (sample mean for topic  $j$ ). Equation (36) generally yields a larger estimate, because while one-way ANOVA removes only the *between-system* variation from the total variation (see Eq. 14), two-way ANOVA without replication removes the *between-topic* variation as well. As we prefer to “err on the side of oversampling” as recommended by Ellis (2010), we use Eq. (36) in this study. Researchers who are interested in tighter estimates are welcome to try our Excel files with their own variance estimates.

As we shall explain in Sect. 4, we have two different topic-by-run matrices (i.e., test collections and runs) for each evaluation measure for every IR task that we consider. To enhance the reliability of our variance estimates, we first obtain a variance estimate  $\hat{\sigma}_C^2$  from each matrix using Eq. (36), and then *pool* the two estimates using the following standard formula<sup>26</sup>:

$$\hat{\sigma}^2 = \frac{\sum_C (n_C - 1) \hat{\sigma}_C^2}{\sum_C (n_C - 1)} \tag{38}$$

As for  $\sigma_i^2 = \sigma_X^2 + \sigma_Y^2$ , we introduce the homoscedasticity assumption here as well and let  $\hat{\sigma}_i^2 = 2\hat{\sigma}^2$ . While this probably overestimates the variances of the score *differences*, again, we choose to “err on the side of oversampling” (Ellis 2010) in this study.

## 4 Data

Table 3 provides some statistics of the past data that we used for obtaining  $\hat{\sigma}^2$ 's. We considered three IR *tasks*: (a) adhoc news search; (b) adhoc web search; and (c) diversified web search; for each task, we used two data sets to obtain pooled variance estimates.

The adhoc/news data sets are from the TREC robust tracks, with “new” topics from each year (Voorhees 2004, 2005). The “old” topics from the robust tracks are not good for our experiments for two reasons. First, the relevance assessments for the old topics were constructed based on old TREC adhoc runs, not the new robust track runs. This prevents us from studying the tradeoff between topic set sizes and pool depths (see Sect. 6.3). Second, the relevance assessments for the old topics are binary, which prevents us from studying the benefit of various evaluation measures that can utilise graded relevance assessments.

The web data sets are from the adhoc and diversity tasks of the TREC web tracks (Clarke et al. 2012, 2013). Note that these diversity data sets have per-intent graded relevance assessments, although they were treated as binary in the official evaluations at TREC.

<sup>25</sup> The sample system mean discussed in Sect. 3.2 was for a future system evaluated over  $n$  topics; the one discussed here is for an existing system evaluated over the  $n_C$  topics of an existing collection.

<sup>26</sup> *Pooled variance* is a technical term in statistics, not to be confused with document pools in IR.

**Table 3** TREC test collections and runs used for estimating  $\sigma^2$ 

Short name	Track	Topics $n_C$	Runs $m_C$	Pool depth	Relevance levels	Documents
(a) Task: adhoc/news						
TREC03new	2003 robust	50 (601–650)	78	125	L0–L2	528,155 (disks 4+5 minus the congressional record)
TREC04new	2004 robust	49 (651–700 minus 672)	78*	100	L0–L2	
(b) Task: adhoc/web						
TREC11w	2011 web–adhoc	50 (101–150)	37	25	L0–L3	Approx. one billion (clueweb09)
TREC12w	2011 web–adhoc	50 (151–200)	28	20/30	L0–L4	
(c) Task: diversity/web						
TREC11wD	2011 web–diversity	50 (101–150)	25	25	L0–L3 per intent	Approx. one billion (clueweb09)
TREC12wD	2011 web–diversity	50 (151–200)	20	20/30	L0–L4 per intent	

The web track relevance grades (Clarke et al. 2012, 2013) were mapped to our relevance levels as follows:  $-2$  and  $0 \rightarrow L0$  (i.e., nonrelevant);  $1 \rightarrow L1$ ;  $2 \rightarrow L2$ ;  $3 \rightarrow L3$ ;  $4 \rightarrow L4$  \* TREC 2004 description-only runs excluded; the same set of runs was used in Webber et al. (2008b)

## 5 Measures

When computing evaluation measures, the usual *measurement depth* (i.e., document cut-off) for the adhoc/news task is  $md = 1000$ ; we considered  $md = 10$  in addition for consistency with the web track tasks. Whereas, we consider  $md = 10$  only for the web tasks as we are interested in the *first* search engine result page.

Table 4 provides some information on the evaluation measures that were used in the present study. For the adhoc/news and adhoc/web tasks, we consider the binary *Average Precision* (AP), *Q-measure* (Q) (Sakai 2005), *normalised Discounted Cumulative Gain* (nDCG) (Järvelin and Kekäläinen 2002) and *normalised Expected Reciprocal Rank* (nERR) (Chapelle et al. 2011), all computed using the NTCIREVAL toolkit.<sup>27</sup> For computing AP and Q, we follow Sakai and Song (2011) and divide by  $\min(md, R)$  rather than by  $R$  in order to properly handle small measurement depths.

For the diversity/web task, we consider  $\alpha$ -nDCG (Clarke et al. 2009) and *Intent-Aware nERR* (nERR-IA) (Chapelle et al. 2011) computed using `ndeval`,<sup>28</sup> as well as *D-nDCG* and *D<sub>#</sub>-nDCG* (Sakai and Song 2011) computed using NTCIREVAL. When using NTCIREVAL, the gain value for each LX-relevant document was set to  $g(r) = 2^x - 1$ : for example, the gain for an L3-relevant document is 7, while that for an L1-relevant document is 1. As for `ndeval`, the default settings were used: this program ignores per-intent graded relevance levels.

<sup>27</sup> <http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html>.

<sup>28</sup> <http://trec.nist.gov/data/web/12/ndeval.c>.

**Table 4** Evaluation measures used in this study

Task type	Measure	Used in tasks such as	Tool
Adhoc	AP	TREC adhoc/robust	NTCIREVAL
	Q	NTCIR CLIR/IR4QA/GeoTime	NTCIREVAL
	nDCG	TREC web adhoc	NTCIREVAL
	nERR	TREC web adhoc	NTCIREVAL
Diversity	$\alpha$ -nDCG	TREC web diversity	ndeval
	nERR-IA	TREC web diversity	ndeval
	D-nDCG	NTCIR INTENT	NTCIREVAL
	D $^\#$ -nDCG	NTCIR INTENT	NTCIREVAL

We refer the reader to Sakai (2014c) as a single source for mathematical definitions of the above evaluation measures.

## 6 Results and discussions

### 6.1 Variance estimates

Table 5 shows the within-system variance estimates  $\hat{\sigma}^2$  that we obtained for each evaluation measure with each topic-by-run matrix. For example, with TREC03new and TREC04new,  $\hat{\sigma}^2 = .0479$  and  $.0462$  according to Eq. (36), respectively, and the pooled variance obtained from these two data sets using Eq. (38) is  $\hat{\sigma}^2 = .0471$  as shown in bold. Throughout this paper, we use these pooled variances for topic set size design: note that the variance estimates are similar across the two data sets for each IR task (a1), (a2), (b), and (c), which suggests that given an existing test collection for a particular IR task, it is not difficult to obtain a good estimate of the within-system variance for a particular evaluation measure for the purpose of topic set size design for a new test collection *for the same task*. The estimates look reliable especially for tasks (a1) and (a2), i.e., adhoc/news, where we have as many as  $m_C = 78$  runs.

It is less clear, on the other hand, whether a variance estimate from one task can be regarded as a reliable variance estimate for the topic set size design of a *different* task. The pooled variance estimate for AP at  $md = 10$  obtained from our *adhoc/news* data is  $.0835$ ; this would be a highly accurate estimate if it is used for constructing an *adhoc/web* test collection, since its actual pooled variance for AP is  $.0824$ . However, the variance estimates for Q, nDCG and nERR are not as similar across tasks (a2) and (b). Hence, if a variance estimate from an existing task is to be used for the topic set size design of a new task, it would probably be wise to choose one of the larger variances observed by considering several popular evaluation measures such as AP and nDCG. In particular, note that variances for the *diversity* measures such as the ones shown in Table 5(c) cannot be obtained from past *adhoc* data that lack *per-intent* relevance assessments: in such a case, using a variance estimate of an evaluation measure that is not designed for diversified search is inevitable. For example, if we know from the TREC11w (i.e., TREC 2011) adhoc/web task experience that the variances are in the  $.0477$ – $.1006$  range as shown in



**Table 5**  $\hat{\sigma}^2$  obtained from the topic-by-run score matrices

	#Runs ( $m_C$ )	#Topics ( $n_C$ )	(I) Raw scores			
(a1) Adhoc/news ( $md = 1000$ )			AP	Q	nDCG	nERR
TREC03new	78	50	.0479	.0471	<u>.0456</u>	.1140
TREC04new	78	49	.0462	.0459	<u>.0456</u>	.1150
Pooled	–	–	<b>.0471</b>	<b>.0465</b>	<b><u>.0456</u></b>	<b>.1145</b>
(a2) Adhoc/news ( $md = 10$ )			AP	Q	nDCG	nERR
TREC03new	78	50	.0894	<u>.0659</u>	.0726	.1205
TREC04new	78	49	.0774	<u>.0630</u>	.0732	.1208
Pooled	–	–	<b>.0835</b>	<b><u>.0645</u></b>	<b>.0729</b>	<b>.1206</b>
(b) Adhoc/web ( $md = 10$ )			AP	Q	nDCG	nERR
TREC11w	37	50	.0877	<u>.0477</u>	.0540	.1006
TREC12w	28	50	.0770	<u>.0258</u>	.0341	.0720
Pooled	–	–	<b>.0824</b>	<b><u>.0368</u></b>	<b>.0441</b>	<b>.0863</b>
(c) Diversity/web ( $md = 10$ )			$\alpha$ -	nERR-	D-	D $\#$ -
			nDCG	IA	nDCG	nDCG
TREC11wD	25	50	.0829	.0885	<u>.0379</u>	.0578
TREC12wD	20	50	.0729	.0798	<u>.0301</u>	.0429
Pooled	–	–	<b>.0779</b>	<b>.0842</b>	<b><u>.0340</u></b>	<b>.0504</b>

Table 5(b), then we could let  $\hat{\sigma}^2 = .1006$  (i.e., the estimate for the unstable nERR<sup>29</sup>) for the topic set size design of a new diversity/web task at TREC 2012. As the actual pooled variances for the TREC12wD task is in the .0301–.0798 range, our choice of  $\hat{\sigma}^2$  would have overestimated the required topic set size for TREC12wD, which we regard as far better than underestimating it and thereby not meeting the set of statistical requirements.

## 6.2 Topic set sizes based on the three methods

In this section, we discuss how the pooled variance estimates shown in Table 5 translates to the actual topic set sizes, using the aforementioned three Excel tools. For the  $t$  test and ANOVA-based topic set size design methods, we only present results under Cohen's five-eighty convention [i.e.,  $(\alpha, \beta) = (.05, .20)$ ] throughout this paper; the interested reader can easily obtain results for other settings by using our Excel tools.

The left half of Table 6 shows the  $t$  test-based topic set size design results under Cohen's five-eighty convention for different minimum detectable differences  $minD$ ; similarly, the right half shows the ANOVA-based topic set size design results with  $m = 2$  for different minimum detectable ranges  $minD$ . Throughout this paper, the smallest topic set size within the same set of statical requirements is underlined.

<sup>29</sup> Expected Reciprocal Rank is known to be an unstable measure because of its *diminishing return* property (Chapelle et al. 2011): every time a relevant document is found in the ranked list, the value of the next relevant document in the list is discounted. While this user model is intuitive, it makes the measure unstable as this means that it relies on only a few data points, i.e., highly ranked relevant documents (Sakai 2014c).

**Table 6** Topic set size table:  $t$  test-based versus ANOVA with  $m = 2$  systems  $[(\alpha, \beta) = (.05, .20)]$

$minD_t$	$t$ test-based	$minD$	ANOVA-based ( $m = 2$ )
(a1) Adhoc/news, $md = 1000$ (AP/Q/nDCG/nERR)			
.05	298/294/289/721	.05	289/286/280/702
.10	76/75/74/182	.10	73/72/71/176
.15	35/35/34/82	.15	33/33/32/79
.20	21/21/20/47	.20	19/19/18/45
.25	14/14/14/31	.25	12/12/12/29
(a2) Adhoc/news, $md = 10$ (AP/Q/nDCG/nERR)			
.05	527/407/460/760	.05	512/396/447/739
.10	134/104/117/192	.10	129/100/112/185
.15	61/47/53/87	.15	58/45/50/83
.20	35/28/31/50	.20	33/26/29/47
.25	23/19/21/33	.25	21/17/19/30
(b) Adhoc/web, $md = 10$ (AP/Q/nDCG/nERR)			
.05	520/233/279/544	.05	505/226/271/529
.10	132/60/72/138	.10	127/57/68/133
.15	60/28/33/63	.15	57/26/31/60
.20	35/17/20/36	.20	32/15/18/34
.25	23/12/14/24	.25	21/10/12/22
(c) Diversity/web, $md = 10$ ( $\alpha$ -nDCG/nERR-IA/D-nDCG/D $_{\#}$ -nDCG)			
.05	492/531/216/319	.05	478/516/209/309
.10	125/135/56/82	.10	120/130/53/78
.15	57/61/26/38	.15	54/58/24/35
.20	33/36/16/22	.20	31/33/14/20
.25	22/24/11/15	.20	20/22/9/13

Note that, when  $m = 2$  (i.e., there are only two systems to compare),  $minD$  (i.e., the minimum detectable difference between the best and the worst systems) reduces to  $minD_t$  of the  $t$  test. It can be observed that the  $t$  test-based and ANOVA-based ( $m = 2$ ) results are indeed very similar. In fact, since one-way ANOVA for  $m = 2$  is equivalent to the *unpaired* (i.e., two-sample)  $t$  test, one would expect the topic set sizes based on the paired  $t$  test to be a little smaller than those based on ANOVA for  $m = 2$  systems, as the former utilises the fact that the two score vectors are paired. On the contrary, Table 6 shows that the  $t$  test-based topic set sizes are slightly larger. This is probably because of the way we obtain  $\hat{\sigma}_t^2$  for the  $t$  test-based design: since we let  $\hat{\sigma}_t^2 = 2\hat{\sigma}^2$  (see Sect. 3.4), if our  $\hat{\sigma}^2$  for the ANOVA-based design is an overestimate, then the error is doubled for  $\hat{\sigma}_t^2$ . Since the topic set size for a paired  $t$  test should really be bounded above by that for the unpaired  $t$  test under the same statistical requirements, we recommend IR researchers to use our ANOVA-based tool with  $m = 2$  if they want to conduct topic set size design based on the paired  $t$  test. While our  $t$  test tool can handle arbitrary combinations of  $(\alpha, \beta)$  unlike the ANOVA-based counterpart, it is unlikely for researchers to consider cases other than  $\alpha = .01, .05, \beta = .10, .20$  in practice. Our ANOVA-based tool can handle all four combinations of these Type I and Type II error probabilities (see Sect. 3.2).

The ANOVA-based results in Table 6(a1) show that if we want to ensure Cohen’s five-eighths convention for a minimum detectable difference of  $minD_t = 0.10$  in AP for an

**Table 7** Topic set size table: CI-based versus ANOVA with  $m = 10, 100$  systems ( $(\alpha, \beta) = (.05, .20)$ )

$\delta$	CI-based	$minD$	ANOVA-based ( $m = 10$ )	ANOVA-based ( $m = 100$ )
(a1) Adhoc/news, $md = 1000$ (AP/Q/nDCG/nERR)				
.05	<b>588/580/ 569/1427</b>	.05	588/580/569/1427	1520/1501/1472/3695
.10	147/145/143/357	.10	148/146/143/357	381/376/369/924
.15	67/66/65/159	.15	66/65/64/159	170/167/164/411
.20	39/38/37/90	.20	38/37/36/90	96/94/93/232
.25	26/25/25/59	.25	24/24/24/58	62/61/60/148
(a2) Adhoc/news, $md = 10$ (AP/Q/nDCG/nERR)				
.05	<b>1041/ 804/909/1503</b>	.05	1041/804/909/1503	2695/2082/2353/3892
.10	259/201/226/376	.10	261/202/228/376	674/521/589/973
.15	116/91/102/167	.15	116/90/102/168	300/232/262/433
.20	67/52/58/95	.20	66/51/58/95	169/131/148/244
.25	44/34/38/62	.25	42/33/37/61	108/84/95/156
(b) Adhoc/web, $md = 10$ (AP/Q/nDCG/nERR)				
.05	<b>1027/ 459/550/1076</b>	.05	1027/459/550/1076	2659/1188/1424/2785
.10	256/116/138/268	.10	257/115/138/270	665/298/356/697
.15	115/53/63/120	.15	115/52/62/120	296/133/159/310
.20	66/31/36/69	.20	65/29/35/68	167/75/90/175
.25	43/21/24/45	.25	42/19/23/44	107/48/58/112
(c) Diversity/web, $md = 10$ ( $\alpha$ -nDCG/nERR-IA/D-nDCG/D $_{\#}$ -nDCG)				
.05	<b>971/1050/ 424/629</b>	.05	971/1050/424/629	2514/2717/1098/1627
.10	242/261/107/157	.10	243/263/107/158	629/680/275/407
.15	109/117/49/71	.15	109/117/48/71	280/303/123/181
.20	62/67/29/41	.20	61/66/27/40	158/171/69/102
.25	41/44/19/27	.25	40/43/18/26	101/109/45/66

adhoc/news task ( $md = 1000$ ), then we would need 73 topics. Similarly, the ANOVA-based results in Table 6(c) show that if we want to ensure Cohen's five-eighty convention for  $minD_t = 0.15$  in nERR-IA for a diversity/web task, then we would need 58 topics. Hence, existing TREC test collections with 50 topics do not satisfy these statistical requirements. We argue that, through this kind of analysis with previous data, the test collection design for the new round of an existing task should be improved. Note, however, that we are aiming to satisfy a set of statistical requirements for *any* set of systems; our results do *not* mean that existing TREC collections with 50 topics are useless for comparing a particular set of systems.

Table 7 shows the CI-based topic set size design results at  $\alpha = .05$  (i.e., 95 % CI) for different CI widths  $\delta$ ; it also shows the ANOVA-based topic set size design results for different minimum detectable ranges  $minD$  under  $(\alpha, \beta, m) = (.05, .20, 10)$  and  $(\alpha, \beta, m) = (.05, .20, 100)$ .<sup>30</sup> Some topic set sizes could not be computed with our CI-based

<sup>30</sup> The setting  $m = 100$  is not unrealistic. For example, the TREC 2011 Microblog track received 184 runs from 59 participating teams (Ounis et al. 2012).

tool due to a computational limitation in the gamma function in Microsoft Excel<sup>31</sup>; however, we observed that the topic set size required based on the CI-based design with  $\alpha = 0.05$  and  $\delta = c$  is almost the same as the topic set size required based on the ANOVA-based design with  $(\alpha, \beta, m) = (.05, .20, 10)$  and  $\min D = c$ , for any  $c$ . Hence, whenever the CI-based tool failed, we used the ANOVA-based tool instead with  $m = 10$ ; these values are indicated in bold. It can indeed be observed in Table 7 that the CI-based topic set sizes and the ANOVA-based ( $m = 10$ ) results are almost the same. Hence, in practice, researchers who want to conduct topic set size design based on CI-widths can use our ANOVA-based tool instead, by letting  $m = 10$ .

Table 7 also shows that when we increase the number of systems from  $m = 10$  to  $m = 100$ , the required topic set sizes are almost tripled. This suggests that it might be useful for test collection builders to have a rough idea of the number of systems that will be compared at the same time in an experiment.

If we compare across the evaluation measures, we can observe the following from Tables 6 and 7:

- For the adhoc/news tasks at  $md = 1000$ , nDCG requires the smallest number of topics; nERR requires more than twice as many topics as AP, Q and nDCG do;
- For the adhoc/news tasks at  $md = 10$ , Q requires the smallest number of topics; again, nERR requires substantially more topics than AP, Q and nDCG do;
- For the adhoc/web tasks, Q requires the smallest number of topics; AP and nERR require more than twice as many as topics as Q does;
- For the diversity/web tasks, D-nDCG requires the smallest number of topics;  $\alpha$ -nDCG and nERR-IA require more than twice as many topics as D-nDCG does.

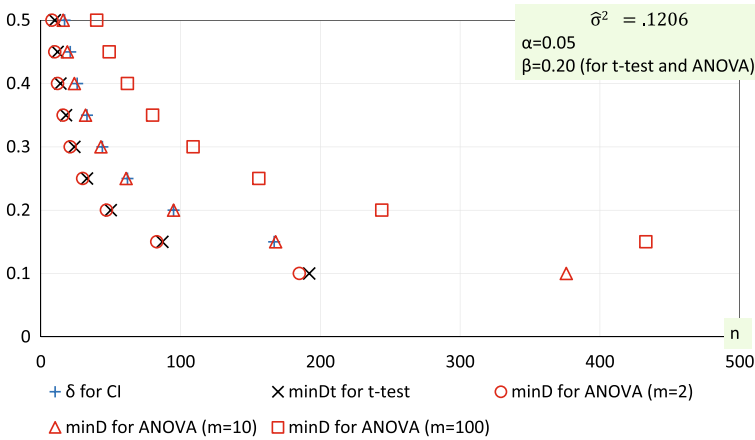
Note that our topic set size design methods thus provide a way to evaluate and compare different measures from a highly practical viewpoint: as the required number of topics is generally proportional to the relevance assessment cost, measures that require fewer topics are clearly more economical. Of course, this is only one aspect of an evaluation measure; whether the measure is actually measuring what we want to measure (e.g., user satisfaction or performance) should be verified separately, but this is beyond the scope of the present study.

Figure 1 visualises the relationships among our topic set design methods. The vertical axis represents the  $\delta$  for the CI-based, the  $\min D_t$  for the  $t$  test-based method, and the  $\min D$  for the ANOVA-based method; the horizontal axis represents the number of topics  $n$ . We used the largest  $\hat{\sigma}^2$  in Table 5, namely,  $\hat{\sigma}^2 = .1206$  for this analysis, but other values of  $\hat{\sigma}^2$  would just change the scale of the horizontal axis. As was discussed earlier, it can be observed that the  $t$  test-based results and the ANOVA-based results with  $m = 2$  are very similar, and that the CI-based results and the ANOVA-based results with  $m = 10$  are almost identical. Also, by comparing the three curves for the ANOVA-based method, we can see how  $n$  grows with  $m$  for a given value of  $\min D$ .

### 6.3 Trade-off between topic set sizes and pool depths for the adhoc/news task

Our discussions so far covered adhoc/news, adhoc/web and diversity/web tasks, but assumed that the pool depth was a given. In this section, we focus our attention on the adhoc/news task (with  $md = 1000$ ), where we have depth-100 and depth-125 pools (see Table 3), which gives us the option of reducing the pool depth. Hence we can discuss the

<sup>31</sup> GAMMA(172) is greater than  $10^{307}$  and cannot be computed by Excel.



**Fig. 1** Effect of  $\delta$ ,  $minD_t$  and  $minD$  on topic set sizes

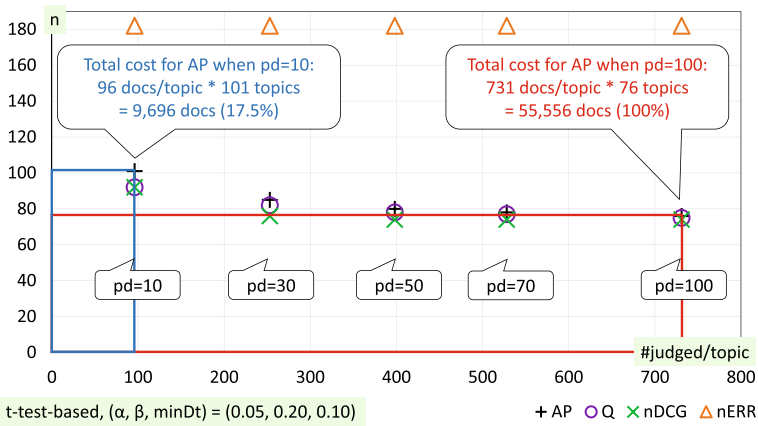
total assessment cost by multiplying  $n$  by the average number of documents that need to be judged per topic for a given pool depth  $pd$ .

From the original TREC03new and TREC04new relevance assessments, we created depth- $pd$  ( $pd = 100, 90, 70, 50, 30, 10$ ) versions of the relevance assessments by filtering out all topic-document pairs that were not contained in the top  $pd$  documents of any run. Using each set of the depth- $pd$  relevance assessments, we re-evaluated all runs using AP, Q, nDCG and nERR. Then, using these new topic-by-run matrices, new variance estimates were obtained and pooled as described in Sect. 3.4.

Table 8 shows the pooled variance estimates obtained from the depth- $pd$  versions of the TREC03new and TREC04new relevance assessments. It also shows the average number of documents judged per topic for each  $pd$ . For example, while the original depth-125 relevance assessments for TREC03new contain 47,932 topic-document pairs, its depth-100 version has 37,605 pairs across 50 topics; the original TREC04new depth-100 relevance assessments have 34,792 pairs across 49 topics. Hence, on average,  $(37,605 + 34,792)/(50 + 49) = 731$  documents are judged per topic when  $pd = 100$ . Similarly,  $(4905 + 4581)/(50 + 49) = 96$  documents are judged per topic when  $pd = 10$ .

**Table 8** Number of relevance assessments versus pooled  $\hat{\sigma}^2$  for reduced pool depths with adhoc/news ( $md = 1000$ )

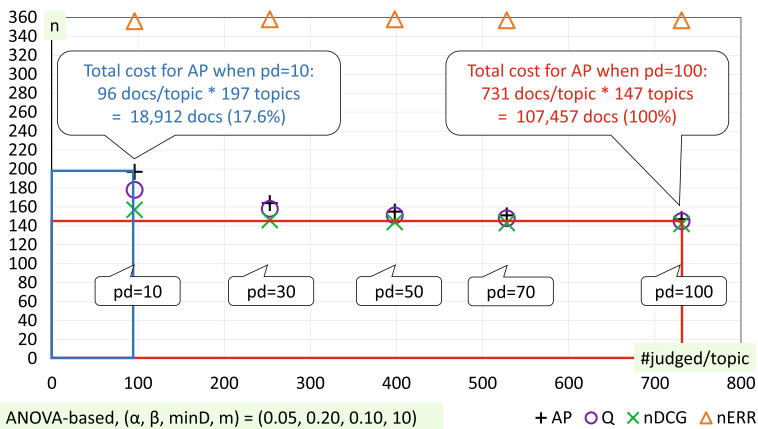
$pd$	TREC03new #judged (50 topics)	TREC04new #judged (49 topics)	Average #judged/topic	Pooled $\hat{\sigma}^2$			
				AP	Q	nDCG	nERR
125	47,932	–	–	–	–	–	–
100	37,605	34,792	731	.0470	.0464	<u>.0454</u>	.1145
70	27,816	24,491	528	.0483	.0474	<u>.0455</u>	.1145
50	20,839	18,612	398	.0494	.0483	<u>.0459</u>	.1146
30	13,045	11,968	253	.0523	.0504	<u>.0467</u>	.1146
10	4905	4581	96	.0630	.0569	<u>.0502</u>	.1141



**Fig. 2** Cost analysis with the *t* test-based topic set size design for the adhoc/news task at *md* = 1000

Based on the *t* test-based method with  $(\alpha, \beta, minD_t) = (.05, .20, .10)$ , Fig. 2 plots the required number of topics *n* against the average number of documents judged per topic for different pool depth settings and different evaluation measures. Recall that the results based on the ANOVA-based method with  $(\alpha, \beta, minD, m) = (.05, 20, .10, 2)$  would look almost identical to this figure. For each pool depth setting, note that the number of topics multiplied by the number of judged documents per topic gives the estimated total assessment cost. Similarly, based on the ANOVA-based method with  $(\alpha, \beta, minD, m) = (.05, 20, .10, 10)$ , Fig. 3 visualises the assessment costs for different pool depth settings. Recall that the results based on the CI-based method with  $\alpha = .05$  would look identical to this figure.

In Fig. 2, the total cost for AP when *pd* = 100 (i.e., the default of TREC adhoc tasks) is 55,556 documents (visualised as the area of a pink rectangle); if we use the *pd* = 10 setting instead, the cost goes down to 9,696 documents (visualised as the area of a blue rectangle). That is, while maintaining the statistical reliability of the test collection, the assessment



**Fig. 3** Cost analysis with the ANOVA-based topic set size design for the adhoc/news task at *md* = 1000

cost can be reduced to  $9,696/55,556 = 17.5\%$ . Similarly, Fig. 3 shows that, if  $m = 10$  systems are to be compared, the assessment cost can be reduced to  $18,912/107,457 = 17.6\%$  by letting  $pd = 10$  instead of the usual  $pd = 100$ . While it is a well-known fact that it is better to have many topics with few judgments per topic than to have few topics with many judgments per topic (e.g., Carterette et al. 2008; Carterette and Smucker 2007; Webber et al. 2008b), our methods visualise this in a straightforward manner.

Figures 2 and 3 also show that because nERR is very unstable, it requires about twice as many topics as the other measures regardless of the choice of  $pd$ . Since the required number of topics is basically a constant for nERR, it would be a waste of assessment effort to construct a depth-100 test collection if the test collection builder plans to use nERR as the primary evaluation measure.<sup>32</sup> Hence, as was discussed earlier, IR test collection builders should probably consider several different evaluation measures at the test collection design phase, take one of the larger variance estimates and plug it into our ANOVA tool, in the hope that the new test collection will meet the set of statistical requirements even for relatively unstable measures. Then, the test collection design  $(n, pd)$  can be re-examined and adjusted after each round of the task.

While the five test collection designs shown in Fig. 2 (and Fig. 3) are statistically equivalent, note that IR test collection builders should collect as many relevance assessments as possible in order to maximise *reusability*, which we define as the ability of a test collection to assess new systems fairly, relative to known systems. That is, if the budget available accommodates  $B$  relevance assessments, test collection builders can first decide on a set of statistical requirements such as  $(\alpha, \beta, \min D, m)$ , obtain several candidate test collection designs  $(n, pd)$  using our ANOVA tool with a large variance estimate  $\hat{\sigma}^2$ , and finally choose the design whose total cost is just below  $B$ .

## 7 Conclusions and future work

In this study, we showed three statistically-motivated methods for determining the number of a new test collection to be built, based on sample size design techniques of Nagata (2003). The  $t$  test-based method and the ANOVA-based method are based on power analysis; the CI-based method requires a tight CI for the difference between any system pair. We pooled the residual variances of ANOVA to estimate the population within-system variance for each IR task and measure, and compared the topic set size design results across the three methods. We argued that, as different evaluation measures can have vastly different within-system variances and hence require substantially different topic set sizes, IR test collection builders should examine several different evaluation measures at the test collection design phase and focus on a high-variance measure for topic set size design. We also demonstrated that obtaining a reliable variance estimate is not difficult for building a new test collection for an existing task, and argued that the design of a new test collection should be improved based on past data from the same task. As for building a test collection for a new task with new measures, we suggest that a high variance estimate from a similar existing task be used for topic set size design (e.g., use a variance estimate from existing adhoc/web task data for designing a new diversity/web task test collection).

<sup>32</sup> Recall that the TREC robust track data used here is for news retrieval (Voorhees 2004), whereas ERR was developed in the context of web search (Chapelle et al. 2011). That is, the observed mismatch between nERR and the depth-100 pools is due to our experimental setting; TREC is not to blame here.

Furthermore, we demonstrated how to study the balance between the topic set size  $n$  and the pool depth  $pd$ , and how to choose the right test collection design ( $n$ ,  $pd$ ) based on the available budget. Our approach thus provides a clear guiding principle for test collection design to the IR community. Note that our approach is also applicable to non-IR tasks as long as a few score matrices equivalent to our topic-by-run matrices are available.

Our Excel tools and the topic-by-run matrices are available online; the interested reader can easily reproduce our results using them with the pooled variance estimates shown in Tables 5 and 8. In practice, since our  $t$  test based results are very similar to our ANOVA-based results with  $m = 2$ , while our CI-based results are almost identical to our ANOVA-based results with  $m = 10$ , we recommend researchers to utilise our ANOVA-based tool regardless of which of our three approaches they want to take.

As for future work, we are currently looking into the use of *score standardisation* (Webber et al. 2008a) for the purpose of topic set size design after removing the topic hardness effect. This requires a whole new set of experiments that involves *leave-one-out* tests (Sakai 2014c; Zobel 1998) in order to study how *new* systems that contributed to neither the pooling nor the setting of per-topic standardisation parameters can be evaluated properly. The results will be reported in a separate study.

While our methods rely on a series of approximations (e.g., Eqs. 5 and 18), these techniques have been compared with exact values and are known to be highly accurate (Nagata 2003). Our view is that the greatest source of error for our topic set size design approach is probably the variance estimation step. Probably the best way to study this effect would be to implement the proposed topic set size design procedure to TREC tracks or NTCIR tasks, update the estimates by pooling the observed variances across the past rounds, and see how the pooled variances fluctuate over time. Our hope is that the variance estimates and the topic set sizes will stabilise after a few rounds, but this has to be verified. We feel optimistic about this, as the actual variances across two rounds of the same track were very similar in our experiments (Table 3).<sup>33</sup> Similarly, we hope to investigate the practical usefulness of our approach for new tasks with new evaluation measures. Can we do any better than just “learn from a similar existing task” as suggested in the present study?

**Acknowledgments** I would like to thank Professor Yasushi Nagata of Waseda University for his valuable advice, and to the guest editors and reviewers for their constructive feedback.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## Appendix 1

Briefly, Nagata (2003) obtained Eq. (5) (a normal approximation of the noncentral  $t$  distribution) as follows.

Let  $\chi^2$  be a random variable that obeys a  $\chi^2$  distribution with  $\phi$  degrees of freedom. The first tool we utilise is the following approximation:

<sup>33</sup> In contrast, Urbano et al. (2013) report that *generalisability theory*, which also relies on variance estimates from past data, is very sensitive to the particular sample of systems and queries used. This discrepancy may also be a worthwhile subject for future research.



$$\sqrt{\frac{\chi^2}{\phi}} \sim N(c^*, \sigma^{*2}), \tag{39}$$

where, using the gamma function, the mean is given by<sup>34</sup>

$$c^* = E\left(\sqrt{\frac{\chi^2}{\phi}}\right) = \frac{E(\sqrt{\chi^2})}{\sqrt{\phi}} = \frac{\sqrt{2}\Gamma\left(\frac{\phi+1}{2}\right)}{\sqrt{\phi}\Gamma\left(\frac{\phi}{2}\right)} \tag{40}$$

and the variance is given by

$$\sigma^{*2} = V\left(\sqrt{\frac{\chi^2}{\phi}}\right) = E\left(\frac{\chi^2}{\phi}\right) - \left(E\left(\sqrt{\frac{\chi^2}{\phi}}\right)\right)^2 = \frac{E(\chi^2)}{\phi} - c^{*2} = 1 - c^{*2} \tag{41}$$

since  $E(\chi^2) = \phi$ .

It is known that, for a random variable  $y$  such that

$$y \sim N(\lambda, 1^2), \tag{42}$$

the following  $t'$  obeys a noncentral  $t$  distribution with  $\phi$  degrees of freedom and a non-centrality parameter  $\lambda$ :

$$t' = \frac{y}{\sqrt{\chi^2/\phi}}. \tag{43}$$

Hence the left hand side of Eq. (5) may be rewritten as:

$$Pr\{t' \leq w\} = Pr\left\{\frac{y}{\sqrt{\chi^2/\phi}} \leq w\right\} = Pr\left\{y - w\sqrt{\frac{\chi^2}{\phi}} \leq 0\right\}. \tag{44}$$

Now, from Eqs. (39) and (42), we obtain:

$$y - w\sqrt{\frac{\chi^2}{\phi}} \sim N(\lambda - wc^*, 1 + w^2\sigma^{*2}). \tag{45}$$

Hence, using a random variable  $u$  that obeys  $N(0, 1^2)$ , Eq. (43) can be rewritten as:

$$Pr\{t' \leq w\} \approx Pr\left\{u \leq \frac{wc^* - \lambda}{\sqrt{1 + w^2\sigma^{*2}}}\right\}. \tag{46}$$

Finally, by using the following approximations for  $c^*$  and  $\sigma^{*2}$ , Eq. (5) is obtained.<sup>35</sup>

$$c^* \approx 1 - \frac{1}{4\phi}, \quad \sigma^{*2} \approx \frac{1}{2\phi}. \tag{47}$$

<sup>34</sup> It is known that  $E((\chi^2)^k) = \frac{2^k\Gamma(\frac{\phi}{2}+k)}{\Gamma(\frac{\phi}{2})}$  holds for  $k > -\phi/2$ .

<sup>35</sup> Johnson and Welch (1940) employ a rougher approximation:  $c^* \approx 1$ ,  $\sigma^{*2} \approx \frac{1}{2\phi}$ .

## Appendix 2

Briefly, Nagata (2003) obtained Eq. (18) (a normal approximation of the noncentral  $F$  distribution) as follows.

Let  $\chi_A^2$  denote a random variable that obeys a *noncentral*  $\chi^2$  distribution (Patnaik 1949) with  $\phi_A$  degrees of freedom and a noncentrality parameter  $\lambda$ ; let  $\chi_E^2$  denote a random variable that obeys a (central)  $\chi^2$  distribution with  $\phi_E$  degrees of freedom. Then, by definition, the following  $F'$  obeys a noncentral  $F$  distribution with  $(\phi_A, \phi_E)$  degrees of freedom and a noncentrality parameter  $\lambda$ :

$$F' = \frac{\chi_A^2/\phi_A}{\chi_E^2/\phi_E} \tag{48}$$

According to Patnaik (1949),  $\chi_A^2/c_A$  can be approximated by  $\chi_A^{*2}$  that obeys a (central)  $\chi^2$  distribution with  $\phi_A^*$  degrees of freedom, where:

$$c_A = \frac{\phi_A + 2\lambda}{\phi_A + \lambda}, \quad \phi_A^* = \frac{(\phi_A + \lambda)^2}{\phi_A + 2\lambda} \tag{49}$$

Therefore, the left hand side of Eq. (18) may be rewritten as:

$$Pr\{F' \leq w\} = Pr\left\{\frac{\chi_A^2/\phi_A}{\chi_E^2/\phi_E} \leq w\right\} \approx Pr\left\{\frac{c_A \chi_A^{*2}}{\phi_A} \leq \frac{w \chi_E^2}{\phi_E}\right\} \tag{50}$$

Using a well-known approximation of a  $\chi^2$  distribution with a normal distribution provided by Fisher (1922), we can assume that:

$$\sqrt{2\chi_A^{*2}} \sim N\left(\sqrt{2\phi_A^* - 1}, 1^2\right), \quad \sqrt{2\chi_E^2} \sim N\left(\sqrt{2\phi_E - 1}, 1^2\right) \tag{51}$$

To utilise the above, let us transform Eq. (50), while preserving the inequality, as follows:

$$\begin{aligned} Pr\{F' \leq w\} &= Pr\left\{\sqrt{\frac{c_A}{\phi_A}}\sqrt{2\chi_A^{*2}} \leq \sqrt{\frac{w}{\phi_E}}\sqrt{2\chi_E^2}\right\} \\ &= Pr\left\{\sqrt{\frac{c_A}{\phi_A}}\sqrt{2\chi_A^{*2}} - \sqrt{\frac{w}{\phi_E}}\sqrt{2\chi_E^2} \leq 0\right\}. \end{aligned} \tag{52}$$

whereas, from Eq. (51), we can assume that:

$$\begin{aligned} &\sqrt{\frac{c_A}{\phi_A}}\sqrt{2\chi_A^{*2}} - \sqrt{\frac{w}{\phi_E}}\sqrt{2\chi_E^2} \\ &\sim N\left(\sqrt{\frac{c_A}{\phi_A}}\sqrt{2\phi_A^* - 1} - \sqrt{\frac{w}{\phi_E}}\left(\sqrt{2\phi_E - 1}, \frac{c_A}{\phi_A} + \frac{w}{\phi_E}\right)\right). \end{aligned} \tag{53}$$

Therefore, using a random variable  $u$  that obeys  $N(0, 1^2)$ , Eq. (18) is obtained. Note that, by letting  $\phi_A = m - 1$  and  $\lambda = n\Delta$  in Eq. (49), Eq. (19) is also obtained.

## References

- Bodoff, D., & Li, P. (2007). Test theory for assessing IR test collections. In *Proceedings of ACM SIGIR 2007* (pp. 367–374).
- Boytsov, L., Belova, A., & Westfall, P. (2013). Deciding on an adjustment for multiplicity in IR experiments. In *Proceedings of ACM SIGIR 2013* (pp. 403–412).
- Carterette, B. (2011). Model-based inference about IR systems. In *ICTIR 2011 (LNCS 6931)* (pp. 101–112).
- Carterette, B. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*. doi:10.1145/2094072.2094076.
- Carterette, B., & Smucker, M. D. (2007). Hypothesis testing with incomplete relevance judgments. In *Proceedings of ACM CIKM 2007* (pp. 643–652).
- Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., & Allan, J. (2008). Evaluation over thousands of queries. In *Proceedings of ACM SIGIR 2008* (pp. 651–658).
- Chapelle, O., Ji, S., Liao, C., Velipasaoğlu, E., Lai, L., & Wu, S. L. (2011). Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6), 572–592.
- Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., et al. (2009). Novelty and diversity in information retrieval evaluation. In *Proceedings of ACM SIGIR 2008* (pp. 659–666).
- Clarke, C. L. A., Craswell, N., Soboroff, I., & Voorhees, E. M. (2012). Overview of the TREC 2011 web track. In *Proceedings of TREC 2011*.
- Clarke, C. L. A., Craswell, N., & Voorhees, E. M. (2013). Overview of the TREC 2012 web track. In *Proceedings of TREC 2012*.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). London: Lawrence Erlbaum Associates.
- Cousineau, D., & Laurencelle, L. (2011). Non-central  $t$  distribution and the power of the  $t$  test: A rejoinder. *Tutorials in Quantitative Methods for Psychology*, 7(1), 1–4.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London: Routledge.
- Ellis, P. D. (2010). *The essential guide to effect sizes*. Cambridge: Cambridge University Press.
- Fisher, R. A. (1922). On the interpretation of  $\chi^2$  from contingency tables and calculation of  $p$ . *Journal of the Royal Statistical Society, Series A*, 85, 87–94.
- Gilbert, H., & Jones, K. S. (1979). Statistical bases of relevance assessment for the ‘IDEAL’ information retrieval test collection. In *Tech. rep., Computer Laboratory, University of Cambridge, British Library Research and Development Report No. 5481*.
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4), 422–446.
- Johnson, N. L., & Welch, B. L. (1940). Applications of the non-central  $t$ -distribution. *Biometrika*, 31, 362–389.
- Jones, K. S., & Van Rijsbergen, C. J. (1975) Report on the need for and provision of an ‘ideal’ information retrieval test collection. *Tech. rep., Computer Laboratory, University of Cambridge, British Library Research and Development Report No. 5266*.
- Jones, K. S., Willet, P., et al. (Eds.). (1997). *Readings in information retrieval*. Los Altos: Morgan Kaufmann.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundation and Trends in Information Retrieval*, 3(1–2), 1–224.
- Killeen, P. R. (2005). An alternative to null hypothesis significance tests. *Psychological Science*, 16, 345–353.
- Nagata, Y. (1998). How to use multiple comparison procedures. *Japanese Journal of Applied Statistics*, 27(2), 93–108. (in Japanese).
- Nagata, Y. (2003). *How to design the sample size*. Tokyo: Asakura Shoten. (in Japanese).
- Nelson, M. J. (1998). Statistical power and effect size in information retrieval experiments. In *Proceedings of CAIS/ASCI’98* (pp. 393–400).
- Okubo, M., & Okada, K. (2012). *Psychological statistics to tell your story: Effect size, confidence interval*. Tokyo: Keiso Shobo. (in Japanese).
- Ounis, I., Macdonald, C., Lin, J., & Soboroff, I. (2012). Overview of the TREC-2011 microblog track. In *Proceedings of TREC 2011*.
- Patnaik, P. B. (1949). The non-central  $\chi^2$ - and  $F$ -distributions and their applications. *Biometrika*, 36, 202–232.

- Sakai, T. (2005). Ranking the NTCIR systems based on multigrade relevance. In *Proceedings of AIRS 2004 (LNCS 3411)* (pp. 251–262).
- Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006* (pp. 525–532).
- Sakai, T. (2014a). Designing test collections for comparing many systems. In *Proceedings of ACM CIKM 2014* (pp. 61–70).
- Sakai, T. (2014b). Designing test collections that provide tight confidence intervals. In *Forum on information technology 2014 RD-003* (Vol. 2, pp.15–18).
- Sakai, T. (2014c). Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between information retrieval and databases (LNCS 8173)* (pp. 116–163).
- Sakai, T. (2014d). Statistical reform in information retrieval? *SIGIR Forum*, 48(1), 3–12.
- Sakai, T. (2014e). Topic set size design with variance estimates from two-way ANOVA. In *Proceedings of EVIA 2014* (pp. 1–8).
- Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of ACM SIGIR 2011* (pp. 1043–1042).
- Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of ACM CIKM 2007* (pp. 623–632).
- Soboroff, I. (2014). Computing confidence intervals for common IR measures. In *Proceedings of EVIA 2014* (pp. 25–28).
- Urbano, J., Marrero, M., & Martín, D. (2013). On the measurement of test collection reliability. In *Proceedings of ACM SIGIR 2013* (pp. 393–402).
- Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.
- Voorhees, E. M. (2004). Overview of the TREC 2003 robust retrieval track. In *Proceedings of TREC 2003*.
- Voorhees, E.M. (2005). Overview of the TREC 2004 robust retrieval track. In *Proceedings of TREC 2004*.
- Voorhees, E. M. (2009). Topic set size redux. In *Proceedings of ACM SIGIR 2009* (pp. 806–807).
- Webber, W., Moffat, A., & Zobel, J. (2008). Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of ACM SIGIR 2008* (pp. 51–58).
- Webber, W., Moffat, A., & Zobel, J. (2008). Statistical power in retrieval experimentation. In *Proceedings of ACM CIKM 2008* (pp. 571–580).
- Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of ACM SIGIR 1998* (pp. 307–314).