TÜBİTAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 2218 – 2232

doi:10.3906/elk-1203-20

Design of information retrieval experiments: the sufficient topic set size for providing an adequate level of confidence

Bekir Taner DİNÇER*
Department of Computer Engineering, Faculty of Engineering, Muğla University, Turkey

Received: 05.03.2012 • **Accepted:** 24.06.2012 • **Published Online:** 30.10.2013 • **Printed:** 25.11.2013

Abstract: In the current design of information retrieval (IR) experiments, a sample of 50 topics is generally agreed to be sufficient in size to perform dependable system evaluations. This article presents the detailed and formal explanation of how the second fundamental theorem of probability, the central limit theorem, can be used for the estimation of the sufficient size of a topic sample. The research performed in this article, using past Text Retrieval Conference data, reveals that, on average, 50 topics will be sufficient to provide a confidence level at or above 95% if the null hypothesis of an equal population mean average precision (MAP) (H_0) is rejected for 2 IR systems having an observed difference in the MAP of 0.035 or more, whereas, in contrast, previous empirical research suggests a difference in the MAP of 0.05 or more. This study also shows that, for individual system pairs, the sample size required to provide 95% confidence on a declared significance may range from a size as small as 10 to a size as large as 722. Thus, for the design of IR experiments, it agrees with the common view that relying on average figures as a rule of thumb may well be misleading.

Key words: Information retrieval system evaluation, topic set size, central limit theorem, generalizability theory

1. Introduction

In the field of information retrieval (IR), system evaluation (or experimental evaluation, or batch evaluation) refers to the relative comparison of the effectiveness of IR systems under the same controlled experimental conditions. The ultimate goal of system evaluation is to decide whether one IR system is better in retrieval effectiveness than the other on the population of information needs or topics. A design paradigm for system evaluation was first introduced in the Cranfield II experiment [1], where IR systems were evaluated using a test collection with 3 fundamental components: a set of documents, a set of posed information needs, and a set of relevance judgments. Relevance judgments are the collections of documents that should be retrieved for each set of information needs, and a posed information need is a query that may be formulated by any inquirer (user). This experimental design paradigm has been in use for over 2 decades, and it is still actively used in almost all large-scale experimental evaluation efforts.

In this paradigm, relevance is the sole effectiveness criterion, and the effectiveness of a system based on relevance is measured in 2 dimensions: the ability to retrieve documents that are known as relevant, and the ability to suppress documents that are known to be nonrelevant. The majority of the currently used measures of relevance are based on precision and recall. Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved.

In the traditional evaluation of retrieval experiments, performances of the systems are measured over a

*Correspondence: dtaner@mu.edu.tr

set of topics (in the Text Retrieval Conference (TREC), the terms 'topic' and 'run' are used to indicate the 'information need' and 'IR system/retrieval strategy', respectively). Since a performance summary measure is necessary to compare different IR retrieval systems over all of the predefined information needs, a final summary performance score for each system is calculated as the average of its performance scores observed on all of the topics. In particular, the mean average precision (MAP) is the most widely used summary measure. A MAP score of a particular system is the mean of the uninterpolated average precisions (APs) observed on all topics, and, in turn, an AP score of a document set retrieved by a system is the average of all of the precision scores that are calculated at each relevant document reached from the start in that document set.

As in the case of the population of topics, populations are usually infinite in size and unknown in distribution. This is the reason why we need to use statistical hypothesis tests for making inductive inferences from samples to a population characteristic of interest. In particular, the validity of an inductive inference depends on the accuracy of estimates in estimating the true value of the population characteristic of interest. On the other hand, the accuracy of an estimate that is derived from a particular sample depends on whether the sample in use is a true representative of the population that we intended to infer to. Thus, to reliably decide whether one IR system is better than the other in the population of topics, we need to estimate the true population effectiveness of the individual IR systems with enough accuracy. On this account, the theory of probability sampling [2] rules the selection of individual observations for the purpose of statistical inference. According to the theory of probability sampling, an estimate that is derived from a random sample is empirically the best estimate of the true value of the population characteristic of interest, with a measurable amount of (sampling) error or uncertainty [3].

Sparck Jones [4] stated that "a difference in scores that is greater than 0.05 is noticeable, and a difference that is greater than 0.10 is material". In the works of Buckley and Voorhees [5] and Voorhees and Buckley [6], the effect of the topic set size on retrieval experiment error rate was investigated, and it was reported, in the latter work, that "an absolute difference in MAP of 0.05–0.06 would be needed between two IR systems measured on 50 topics before concluding, with 95% confidence, that the same systems ranking can be obtained on a different set of 50 topics". In the same line of research, Webber et al. [7] conducted an empirical research based on statistical power analysis and reported that "the standard 50 topics TREC collection can only be relied on to detect true AP deltas [MAP differences] in the range 0.06–0.08". This explains why "a large enough difference between two effectiveness scores" is a generally accepted rule of thumb for performing a dependable system evaluation in the IR community.

Voorhees [8] was the first researcher to perform empirical research on the sufficiency of the TREC standard sample of 50 topics, saying, "at least for the TREC-6 environment, as few as 25 topics can be used to compare the relative effectiveness of different retrieval runs with great confidence". In her more recent work, however, Voorhees [9] recommended the following: "Fifty-topic sets are clearly too small to have confidence in a conclusion when using a measure as unstable as $P(10)$ [precision at 10 documents]. Even for stable measures, researchers should remain skeptical of conclusions demonstrated on only a single test collection".

Note that the previous empirical research tried to single out, once and for all, a lower boundary for the difference of 2 MAP scores, above which every difference can be considered as significant based on a test collection with 50 topics. These past studies suggest, in general, a lower boundary that is not less than 0.05 as a measured MAP. However, note that observing a MAP difference of less than 0.05 is not a rare event in an ordinary IR system evaluation. Thus, it is not unlikely that we need more than 50 topics in practice. The question therefore arises as to "do we necessarily need more than 50 topics for every system pair of between

which the observed MAP difference is less than 0.05?" This is the research question of interest in this article, to which none of the key empirical works have given a precise answer.

Given a particular pair of IR systems, a topic sample size that is sufficient to give significance to the observed MAP difference between the systems may be insufficient to give significance to the same MAP difference if it is observed between another system pair, due to the differences of inherent variability in AP scores across topics. The sufficiency of a topic sample size is subjective to the system pair under consideration. This is a point that most of the previous research primarily overlooked.

The organization of this article is as follows. In the next section, the amount of uncertainty in estimating the population MAP of an IR system is estimated using past TREC 6, 7, and 8 data on the basis of the central limit theorem (CLT). Afterwards, the amount of uncertainty in estimating the difference of the population MAPs of 2 IR systems is estimated in the following section, and the conclusion is given subsequently.

2. Estimation of the population MAP of an IR system

In parametric statistics, it is assumed that a target population can be generated by a well-known distribution, such as normal, exponential, or Poisson, having 1 or more parameters, at least 1 of which is unknown and must be inferred. The population characteristic of interest is usually the unknown parameter or a function of it. A series of independent random variables $X_1, X_2, \ldots, X_n$, each of which has the same distribution on the population, is called an independent and identically distributed (i.i.d.) random sample of size n. An estimator for the population characteristic of interest is formulated as a function of $X_1, X_2, \ldots, X_n$, such that the estimate can be derived from a single, 'observed' sample $x_1, x_2, \ldots, x_n$.

Unfortunately, no inductive inference is certain, so every statement drawn from experimental data is subject to uncertainty. An estimate that is derived from a single sample is subject to uncertainty because of having only 1 sample; that is, different samples from the same population would, in general, yield different estimates. The amount of uncertainty associated with an estimate is inversely proportional to the amount of (population) information contained in the sample that the estimate is derived from. The question therefore arises as to whether it is possible to ascribe a measure of information to the various possible experimental designs available in order to consider the cost of obtaining a particular amount of information: is it worth that cost and at what stage is the cost of obtaining further information too great? Suppose that the purpose of an experiment is to estimate a single parameter of a population distribution. The only requirement that the measure of information should satisfy is that the information on a parameter provided by, say, 2 independent samples drawn from the population should be equal to the sum of the information contained in the 2 samples considered separately. This means that the information contained in a sample should be directly proportional to the sample size n.

The generally adapted measure of information, which was introduced by Fisher [10], is given by:

$$nI = n \int_{-\infty}^{\infty} \left(\frac{\partial \log f}{\partial \theta} \right)^2 f \quad dx,$$

where θ, which we wish to estimate, is the parameter of the population distribution $f(x, \theta)$. For the case where the population parameter of interest is the mean μ of a normal distribution with variance σ^2, the total information contained in a sample is given by the ratio of the sample size n to the population variance σ^2, such that:

$$nI = \frac{n}{\sigma^2}.$$

For the mean of a sample of size n_1 it is n_1/σ^2, and for another sample of size n_2, it is n_2/σ^2, while for the mean of the combined sample it is $(n_1+n_2)/\sigma^2$. This suggests that the information per observation is $1/\sigma^2$ and, most importantly, the information about the mean of a normal distribution is contained entirely in the variance σ^2.

An estimate that is derived from a single sample is subject to uncertainty because of having only 1 sample, but such an estimate necessarily follows a particular distribution on the samples that could be drawn from the same population. In statistics, this distribution is called the sampling distribution (or the null distribution), and it is the measure of uncertainty [3]. In estimating the population mean μ, the amount of uncertainty associated with the mean of a sample of size n ($\bar{x}$) is equal to the variance of the sampling distribution of $\bar{x}$ around μ, σ^2/n, i.e. $1/nI$.

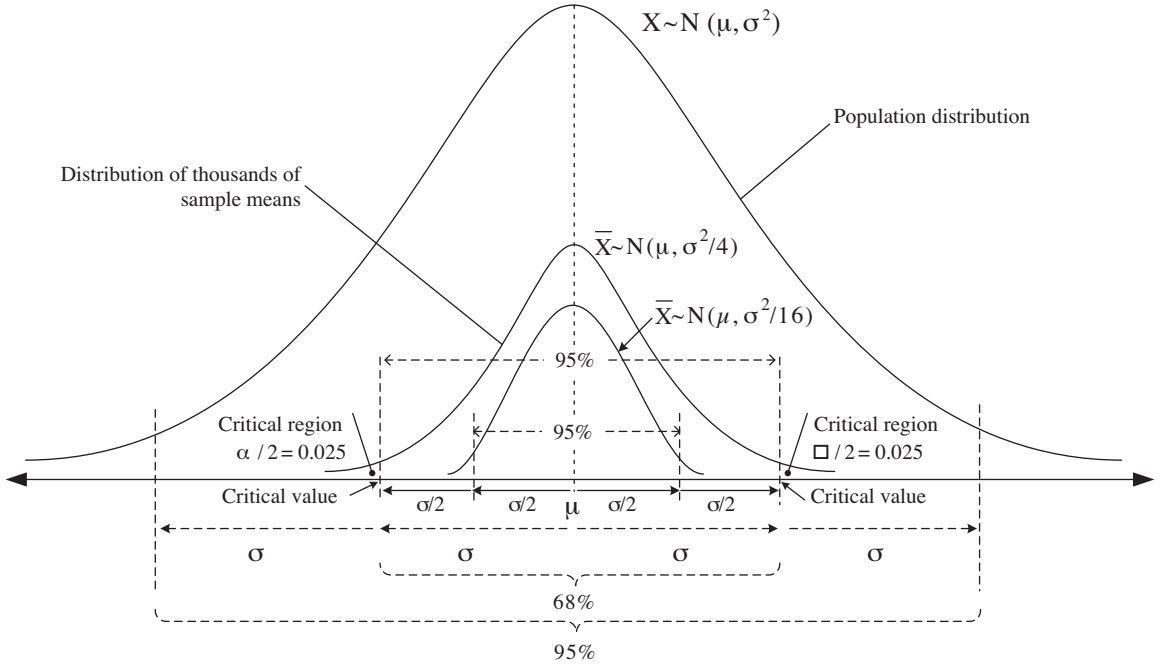

Figure 1. Sampling distribution of the sample mean on the samples of size 4 and 16 from a normal distribution with mean μ and variance σ^2.

Figure 1 illustrates the sampling distributions of the mean of a sample of size $n=4$ and the mean of a sample of size $n=16$ from a normal distribution with mean μ and variance σ^2 (i.e. $X \sim N(\mu,\sigma^2)$).

In the population, the interval $[\mu-\sigma, \mu+\sigma]$ is expected to contain about 68% of all observations and the interval $[\mu-2\sigma, \mu+2\sigma]$ contains about 95% of all observations. On the other hand, for the sampling distributions, it is expected that, 95% of the time, the interval $[\mu-\sigma, \mu+\sigma]$ contains the mean of a sample of 4 observations and the interval $[\mu-\sigma/2, \mu+\sigma/2]$ contains the mean of a sample of 16 observations. The true population mean μ will therefore be with $\bar{x}\pm\sigma$ of the mean of a sample of size $n=4$, and with $\bar{x}\pm\sigma/2$ of the mean of a sample of size $n=16$, 95% of the time. Here, 95% of the time refers to 95% of the samples that could be drawn from the population, and, in turn, 95% of the samples that could be drawn from the population refers to a confidence level of 95% or a significance level of 5% (i.e. $\alpha=0.05$). Note that quadrupling the size of a sample reduces the amount of uncertainty only by a factor of 2. In relation to the IR system evaluations,

the population in Figure 1 can be thought of as the population of topics, where observations represent the AP scores of an IR system and the mean of a sample of size n represents the associated MAP score measured on a sample of n topics.

For a normal distribution with mean μ and variance σ^2, it can be shown that the transformation

$$Z = \frac{\bar{x} - \mu}{SD_{\bar{X}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution with zero mean and unit variance. In other words, for normally distributed populations, $\bar{x}$ follows a normal distribution with mean μ and variance σ^2/n, i.e. $\bar{x} \sim N(\mu, \sigma^2/n)$.

Given a series of n i.i.d. random variables, $X_1, X_2, \ldots, X_n$, each of which follows the same distribution with finite mean μ and variance $\sigma^2 > 0$ (i.e. a 'well-behaved' distribution), the second fundamental theorem of probability, the CLT, assures that as the sample size n increases, the distribution of the sample mean of the random variables weakly converges in probability to a normal distribution with mean μ and variance σ^2/n, irrespective of the shape of the population distribution. Note that since the population parameters are usually not known, σ^2 needs to be estimated from sample statistics. In this regard the well-known estimator of σ^2 is the sample variance s^2, and so $s/\sqrt{n}$ can be used for estimating the standard deviation of the sampling distribution of $\bar{x}$, $SD_{\bar{X}}$. It immediately follows that the distribution of the transformation

$$z = \frac{\bar{x} - \mu}{SE_{\bar{X}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

can be approximated by a standard normal distribution with zero mean and unit variance for any well-behaved population distribution, provided that the sample size n is large enough (generally agreed to be $n \geq 30$).

By elaborating the z transformation given, we can determine the sample size required to provide a desired level of accuracy in estimating μ, such that:

$$n_{\delta,\alpha} \geq \left(\frac{s \cdot z_{\alpha/2}}{\bar{x} - \mu} \right)^2 . \tag{1}$$

Here, $z_{\alpha/2}$ is the z-score, where in between $\pm z_{\alpha/2}$, the area under the standard normal curve is summed up to $1-\alpha$, e.g., $\pm z_{\alpha/2} = \pm 1.96$ at $\alpha = 0.05$. The discrepancy between $\bar{x}$ and μ, $|\bar{x} - \mu|$, is commonly referred to as the sensitivity (or the maximum error, or the error of estimate, the maximum error of estimate, the maximum allowable error, etc.) and is denoted by δ. Sensitivity is the desired accuracy in estimating the true mean of a population distribution, such that $Pr(|\bar{x} - \mu| \geq \delta) \leq \alpha$ or $Pr(|\bar{x} - \mu| \leq \delta) \geq (1 - \alpha)$.

Table 1 lists the topic sample size estimates that are required to maintain the levels of sensitivity $\delta = \pm 0.01, \pm 0.02, \ldots, \pm 0.07$ in estimating the true population MAPs of the first 10 TREC 6 Category-A, automatic, short (Title + Description) runs, with at most, a 5% sampling error. As seen, it is expected, on average, that the true population MAP of a TREC 6 run will be with ± 0.05 of the MAP observed on, at least, 95% of the samples of 50 topics that could be drawn from the population.

The summary statistics for TREC 6, 7, and 8 are given in Table 2. As seen, when the topic sample size $n \geq 50$, it is expected, on average, that $Pr(|\bar{x} - \mu| \leq 0.05) \geq 0.95$, or complementarily $Pr(|\bar{x} - \mu| \geq 0.05) \leq 0.05$.

Table 1. Topic sample sizes suggested by Inequality 1 for maintaining the $\delta = \pm 0.01, \pm 0.02, \ldots, \pm 0.07$ sensitivity in estimating the true population MAPs of the first 10 TREC 6 runs, with at most 5% sampling error or at least 95% confidence. The last 2 blocks of the rows list the averages over the first 10 runs and the total of 29 runs, respectively. The percentage differences associated with the corresponding sensitivity levels relative to 'Average MAP' are listed through '% Diff' rows, e.g., 0.05 / 0.1380 = 36%. The '% Diff from top' row lists the percentage difference relative to the top MAP of 0.2164, e.g., 23% for ±0.05 sensitivity.

				Sensitivity (δ)						
Rank	Runs	MAP	s^2	0.01	0.02	0.03	0.04	**0.05**	0.06	0.07
1	city6ad	0.2164	0.0575	2211	553	246	138	88	61	45
2	LNaShort	0.1972	0.0328	1261	315	140	79	50	35	26
3	uwmt6a2	0.1912	0.0484	1859	465	207	116	74	52	38
4	att97ac	0.1847	0.0445	1711	428	190	107	68	48	35
5	Cor6A2qtcs	0.1809	0.0439	1686	421	187	105	67	47	34
6	att97ae	0.1801	0.0437	1679	420	187	105	67	47	34
7	Cor6A1cls	0.1799	0.0429	1650	412	183	103	66	46	34
8	VrtyAH6a	0.1784	0.0452	1738	435	193	109	70	48	35
9	ibms97a	0.1775	0.0350	1346	337	150	84	54	37	27
10	ibmg97a	0.1727	0.0356	1369	342	152	86	55	38	28
Average		0.1859	0.0430	1651	413	183	103	66	46	34
% Diff				5	11	16	22	27	32	38
Total of 29 runs		0.1380	0.0305	1175	294	131	73	**47**	33	24
% Diff				7	14	22	29	**36**	43	51
% Diff from top				5	9	14	18	**23**	28	32

Table 2. Summary statistics of the topic sample size estimates to maintain $\delta = \pm 0.01, \pm 0.02, \ldots, \pm 0.07$ sensitivity in estimating the true population MAPs of TREC 6, 7, and 8 runs, with at most 5% sampling error or at least 95% confidence.

		Sensitivity (δ)						
Aver. s^2		0.01	0.02	0.03	0.04	**0.05**	0.06	0.07
TREC 6	0.0305	1175	294	131	73	47	33	24
TREC 7	0.0248	953	238	106	60	38	26	19
TREC 8	0.0379	1457	364	162	91	58	40	30
Average	0.0311	1195	299	133	75	**48**	**33**	**24**

3. Estimation of the difference of 2 population MAPs

Let the 2 series of i.i.d. random variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ denote, respectively, the samples of the AP scores of 2 IR systems, A and B, which are measured on a topic sample of size n. Suppose that the random variables X and Y distribute independently and normally on the population of topics; that is, $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are mutually independent in distribution, where μ_X and σ_X^2 denote, respectively, the mean and the variance of the population AP distribution of A, and μ_Y and σ_Y^2 are of B. In a similar manner as σ^2/n gives the uncertainty associated with the mean of a normal sample of size n, the uncertainty associated with $\bar{X} - \bar{Y}$ in estimating $\mu_X - \mu_Y$ is given by:

$$\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}.$$

It can be shown that when $\mu_X = \mu_Y$,

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{(\sigma_X^2 + \sigma_Y^2)/n}} = \frac{\bar{x} - \bar{y}}{SD_{\bar{x}-\bar{y}}}$$

follows a standard normal distribution with zero mean and unit variance, where $SD_{\bar{x}-\bar{y}}$ is the population standard deviation of the difference of 2 sample means, i.e. the standard deviation of the sampling (or rather null) distribution of $\bar{x}-\bar{y}$ around $\mu_X-\mu_Y = 0$. For large sample sizes, the distribution of z can be approximated, under the null hypothesis $H_0 : \mu_X = \mu_Y$, by a standard normal distribution with zero mean and unit variance, irrespective of the shapes of the population AP distributions of A and B, because of the CLT.

The standard deviation of $\bar{x} - \bar{y}$, $SD_{\bar{x}-\bar{y}}$, can be derived from sample statistics in 2 different ways, depending on whether or not the corresponding population distributions are mutually independent.

If 2 population distributions are mutually independent (i.e. the case derived so far), the sample estimate of $SD_{\bar{X}-\bar{Y}}$ is given by:

$$SE_p = \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{n}} = \sqrt{\left(\frac{s_x^2 + s_y^2}{2}\right) \times \frac{2}{n}} = s_p \times \sqrt{\frac{2}{n}},$$

where s_p denotes the pooled standard deviation, and for our case, s_x^2 and s_y^2 denote, respectively, the sample variances of the AP scores of A and B.

On the other hand, when 2 samples (of AP scores) come from 2 dependent population distributions, each of the samples contains a particular amount of information about the other sample due to the fact that each value in 1 population is related or linked to a specific value in the other population; that is, the AP scores of A correlate with the AP scores of B across topics, and vice versa. To utilize the common information contained in the samples, the sample estimate of $SD_{\bar{X}-\bar{Y}}$ is calculated accordingly, such that:

$$SE_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} \times \sqrt{\frac{1}{n}} = s_d \times \sqrt{\frac{1}{n}},$$

where s_d denotes the sample standard deviation of the paired differences $d_i = x_i - y_i$ for $i = 1, 2, \ldots, n$ (i.e. the paired (sample) standard deviation), and $\bar{d} = (1/n) \sum d_i$.

The former $SD_{\bar{X}-\bar{Y}}$ estimate, SE_p, is used for testing the significance of the difference of 2 sample means under the assumption of independence, i.e. the Student t-test for 2 independent sample means. The latter one, SE_d, is used for the same purpose but under the assumption of dependence. This protocol of hypothesis testing is commonly referred to as the Student t-test for 2 dependent sample means or matched pairs, which is in fact the most widely used hypothesis test in the current practice of IR system evaluations.

By elaborating the z transformation given, we can now determine the sample size required to provide $100 \times (1-\alpha)\%$ confidence on a declared significance (i.e. the case of succeeding in rejecting the null hypothesis $H_0 : \mu_X = \mu_Y$) at a predefined level of significance α, such that if the sample size is,

$$n_{\delta,\alpha} \geq \left(\frac{s \cdot z_{\alpha/2}}{\delta}\right)^2, \tag{2}$$

then $Pr(|\bar{x} - \bar{y}| \geq \delta) \leq \alpha$ when $\mu_X = \mu_Y$, so $Pr(|\mu_X - \mu_Y| > 0) \geq (1 - \alpha)$ when $|\bar{x} - \bar{y}| \geq \delta$. Under the assumption of independence, s denotes s_p; otherwise, s_d.

Note that here, Inequality 2 yields the total size of the 2 AP samples, $X_1, X_2, \ldots, X_n$ and Y_1, Y_2, ..., Y_n, not the topic sample size, which is required to provide the amount of population information necessary to maintain a particular level of sensitivity (δ) in estimating the difference of the population MAPs of A and B. This is the same for Inequality 1, but it is implicit; there is only 1 AP sample, $X_1, X_2, \ldots, X_n$, to measure on a sample of n topics. In brief, when X and Y are independent in distribution, the total sample size yielded from Inequality 2 with s_p will be twice that of the topic sample size that is required in effect (as indicated by the $\sqrt{2/n}$ component of the $SD_{\bar{x}-\bar{y}}$ estimate under independence), while, in contrast, the total sample size that is yielded from Inequality 2 with s_d will be equal to the required topic sample size (as indicated by the $\sqrt{1/n}$ component of the $SD_{\bar{x}-\bar{y}}$ estimate under dependence). The consequence is that when the AP scores of 2 IR systems are mutually independent in distribution on the population of topics, the half of the sample size suggested by Inequality 2 with s_p will be equal to the sample size suggested by Inequality 2 with s_d, and, in effect, both will be equal to the required topic sample size.

Without loss of generality, consider the 2 TREC 6 runs in Table 1, 'city6ad' and 'LNaShort'. The associated pooled standard deviation s_p is 0.2125 (i.e. $\sqrt{(0.0575 + 0.0328)/2} = \sqrt{0.0452}$) and the paired standard deviation s_d is 0.1479. Setting δ to, say, 0.05 in Inequality 2 yields a sample size estimate equal to $[(0.2125 \times 1.96)/0.05]^2 \approx 69$ under the assumption of independence and $[(0.1479 \times 1.96)/0.05]^2 \approx 34$ under the assumption of dependence. Here, the fact that $69/2 \approx 34$ suggests that the observed AP scores of 'city6ad' and 'LNaShort' would have arisen from 2 independent distributions on the population of topics. Note that the figure yielded from Inequality 2 with s_p can also be obtained by averaging the individual sample size estimates yielded from Inequality 1 for 'city6ad' and 'LNaShort' (in Table 1), such that $(88 + 50)/2 = 69$.

It appears that, if 'city6ad' and 'LNaShort' have equal MAPs on the population of topics, a MAP difference that is equal to or less than 0.05 would be observed, by chance, on at least 95% of the samples of 34 topics that could be drawn from the topic population. Thus, at least a difference in MAP of 0.05 is needed between 'city6ad' and 'LNaShort' measured on 34 topics before concluding, with 95% confidence, that they do not have equal MAPs on the population of topics (or, equivalently, that the same ranking of 'city6ad' and 'LNaShort' can be observed on a different sample of 34 topics).

The actual difference in the MAPs of 'city6ad' and 'LNaShort', which is observed for the original sample of 50 TREC 6 topics, is 0.0192. A sensitivity level of $\delta = \pm 0.05$ is therefore too low to decide whether there exists a population effect between them. In other words, a difference in MAP of 0.0192 is not unlikely on a sample of 34 topics when $\mu_X = \mu_Y$, so it can be attributed to chance fluctuation. In common practice, δ is usually set to the observed difference $|\bar{x} - \bar{y}|$ (or less), since a sensitivity level greater than the difference observed could have no particular meaning under the null hypothesis. When δ is set to 0.0192, Inequality 2 yields a topic sample size estimate of $[(0.1479 \times 1.96)/0.0192]^2 = 228$. Thus, we can conclude that a sample of 50 topics is insufficient in size to provide the level of sensitivity that could supply the empirical evidence to reject the null hypothesis that 'city6ad' and 'LNaShort' have equal population MAPs.

The current sensitivity level provided by the sample of 50 TREC 6 topics can be estimated by elaborating Inequality 2, as given by:

$$\delta \geq \frac{s \cdot z_{\alpha/2}}{\sqrt{n}} = \frac{0.1479 \times 1.96}{\sqrt{50}} = \pm 0.0409.$$

This quantity expresses the same fact based on MAP difference instead of topic sample size: a difference in MAP of 0.0192 can be attributed to chance fluctuation.

Note that the level of sensitivity at a predefined level of significance α actually determines a $100\times(1-\alpha)\%$ confidence interval that is expected to contain $\mu_X - \mu_Y$ with probability $1-\alpha$:

$$Pr\left[(\bar{x}-\bar{y}) - \frac{s\cdot z_{\alpha/2}}{\sqrt{n}} \leq \mu_X - \mu_Y \leq (\bar{x}-\bar{y}) + \frac{s\cdot z_{\alpha/2}}{\sqrt{n}}\right] \geq 1-\alpha.$$

Thus, when 2 IR systems have equal MAPs on the population of topics, it is expected that the observed MAP difference between the systems will fluctuate by chance across the samples of n topics but in between $\pm z_{\alpha/2}\times s/\sqrt{n}$ with probability $1-\alpha$, such that:

$$Pr\left(-\frac{s\cdot z_{\alpha/2}}{\sqrt{n}} \leq \bar{x}-\bar{y} \leq +\frac{s\cdot z_{\alpha/2}}{\sqrt{n}}\right) \geq 1-\alpha,$$

e.g., $Pr\left(-0.0409 \leq \bar{x}-\bar{y} \leq +0.0409\right) \geq 0.95$.

Given a sample of 50 topics, an observed difference in MAP of 0.0409 or more can therefore be assumed large enough to reject, with 95% confidence, that 'city6ad' and 'LNaShort' have equal population MAPs, or, in other words, to rely on the idea that the probability of rejecting H_0 when it is true will be at or below the nominal error rate of 5%. Note that this is exactly the rationale behind the Student t-test.

As a result, this means that the observed MAP difference between 'city6ad' and 'LNaShort' is not large enough to consider it statistically significant. The statistical analysis performed here is inconclusive in the objective sense due to the lack of enough population information (or empirical evidence). In principle, whenever a statistical analysis is inconclusive, further research should be encouraged before making any conclusions. Nevertheless, one may also accept the null hypothesis of equal population MAPs if further research is not possible, or rather if the current level of sensitivity provided by the sample of topics at hand is so high that the undetectable differences in population MAPs can anyway be considered negligible or unimportant from a practical standpoint.

In theory, by repeating this analysis for every pair of IR systems available and then averaging over all system pairs, we can get a topic sample size estimate that is sufficiently generalizable to be applied to any system pair, on average. However, before making such a generalization, 2 previous key empirical works should be discussed.

First, Lin and Hauptmann [11] showed that there is a resemblance between Fisher's information and the so-called retrieval experiment error rate (REER), which was coined by Voorhees and Buckley [6], such that the model Voorhees and Buckley empirically fitted REER, $b_1\exp[-b_2\cdot n]$, can be approximated by a theoretic model as given by:

$$REER \approx \frac{1}{2}\exp\left[-\frac{2}{\pi}(\mu_X-\mu_Y)^2\frac{n}{(\sigma_X^2+\sigma_Y^2)}\right].$$

This theoretical model suggests that the original REER algorithm assumes independence for the AP distributions of IR systems on the population of topics, as indicated by the following component:

$$\frac{n}{\sigma_X^2+\sigma_Y^2}.$$

Figure 2 illustrates the relationship between Inequality 2 and the REER, based on a topic sample of size $n=4$. As seen in the top panel of Figure 2, if significance is considered to be 2-sided ($\alpha/2=0.025$), it is expected that the sign of the difference between 2 sample means be maintained 97.5% of the time at $\alpha=0.05$ when

the difference of the means of 2 independent population AP distributions is equal to $4\sigma/\sqrt{n} = 4\sigma/2$. This corresponds to a 2.5% REER, as shown in the bottom panel. This means that a topic sample size, which is suggested by Inequality 2 with s_p and $z_{0.025} = 1.96$, would be sufficient to maintain a REER of 2.5% for a pair of 2 independent IR systems having a difference in the population MAP of $4\sigma/\sqrt{n}$. To estimate the topic sample size required to maintain a REER of 5% at $\alpha = 0.05$ by means of Inequality 2, we therefore need to consider the significance as 1-sided (i.e. $z_{0.05} = 1.64$), as illustrated in the bottom panel of Figure 2, i.e. the 'swap area'.

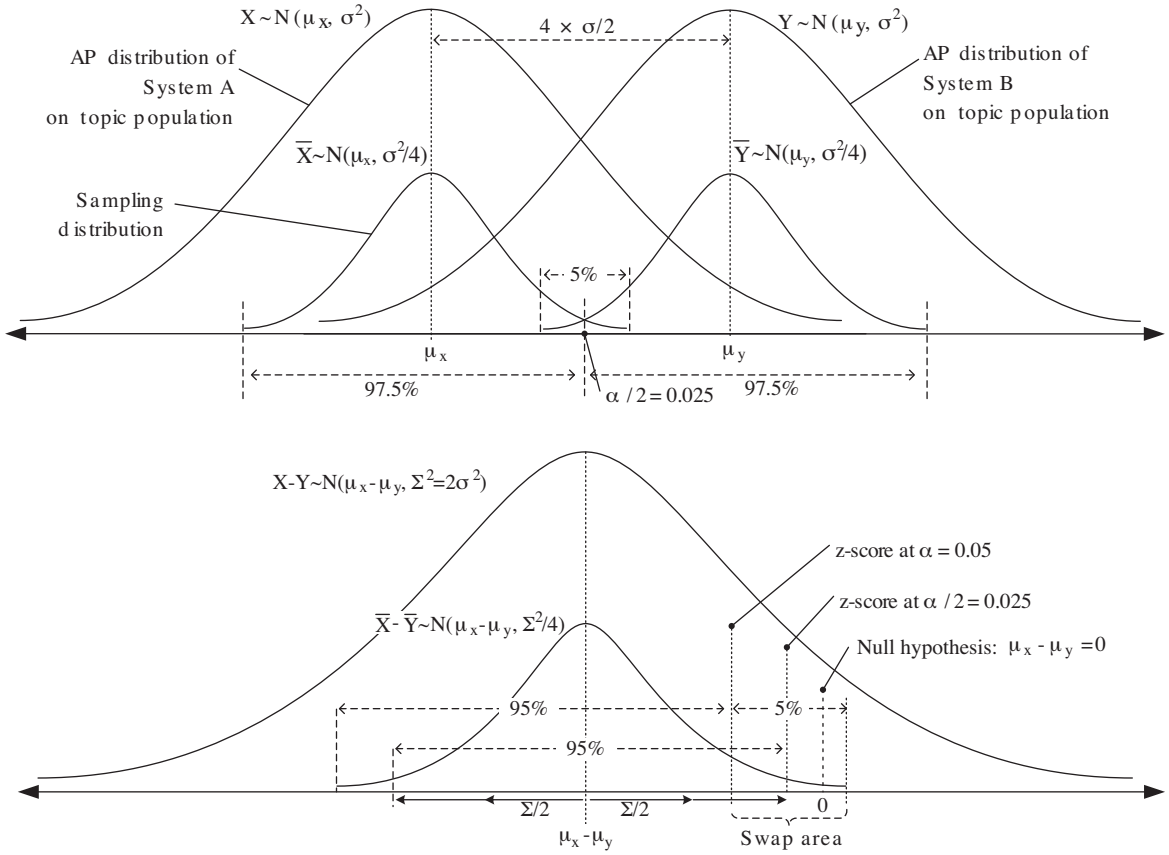

Figure 2. Illustration of 2 AP distributions having a difference in the mean of $4 \times \sigma/\sqrt{n}$ and an equal variance on the population of topics (top panel). $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$ are independent in distribution. The sampling distributions $\bar{X} \sim N(\mu_X, \sigma^2/n)$ and $\bar{Y} \sim N(\mu_Y, \sigma^2/n)$ each represent a sensitivity level of $\delta = \pm 2 \times \sigma/\sqrt{n}$ in estimating the corresponding population means at $\alpha = 0.05$ (assuming $z_{0.025} \approx 2$), such that $|\bar{x} - \mu_X| \leq \delta$ and $|\bar{y} - \mu_Y| \leq \delta$. The bottom panel illustrates the corresponding distribution of the difference of $\bar{x}$ and $\bar{y}$ both for 2-sided significance (z-score at $\alpha/2$) and 1-sided significance (z-score at α).

The research question addressed in the work of Voorhees and Buckley [6] may be restated as "what percentage of the time is the consistency of a size of difference between $\bar{x}$ and $\bar{y}$ maintained in sign across the topic samples of a given size n?" Denote $\bar{x} - \bar{y}$ by D. Next, the original REER algorithm can be viewed as a (nonparametric) resampling technique by means of which one can estimate the sampling distribution of D for any given sample size $n > 0$. It can be shown that, when $\mu_X = \mu_Y$, a sample size that is greater than 0 ($n > 0$) maintains the consistency of every size of difference between $\bar{x}$ and $\bar{y}$ in sign 50% of the time, or rather

50% of the samples of topics that could be drawn from the topic population. More precisely, when $\mu_X = \mu_Y$, $Pr(\bar{x} - \bar{y} \leq 0) = Pr(\bar{x} - \bar{y} \geq 0) = 0.50$.

A sampling distribution that is centered on $\mu_X - \mu_Y = 0$ corresponds to a 50% REER. For 2 IR systems, if the estimated sampling distribution is centered on 0, we can conclude that the observed AP scores of the systems would have arisen from 2 population AP distributions with equal means. On the other hand, if $\mu_X - \mu_Y = 0$ falls within 1 of the 2 critical regions of the estimated sampling distribution, we can reject, at this time, the null hypothesis with, say, 95% confidence at $\alpha = 0.05$, as is shown in the bottom panel of Figure 2. Here, the ratio of the number of samples on which $D \geq 0$ (or $D \leq 0$) to the total number of samples simulated gives the REER associated with a given pair of systems, i.e. the P-value, or the probability, under the null hypothesis, of observing a size of difference in MAP between 2 IR systems as extreme as or more extreme than that observed for the sample of topics given. This means that, over the pairs of runs available in each TREC, by averaging the topic sample size estimates that are yielded from Inequality 2 with s_p and $z_{0.05} = 1.64$ (i.e. 1-sided significance), we can obtain theoretic REER estimates to which the empirical estimates yielded from the original REER algorithm in [9] are expected to converge as the number of run pairs goes to infinity.

Second, Bodoff and Li [12] recently proposed a generalizability theory for the optimization of the design of IR experiments. Although it is not mentioned in the original work of Bodoff and Li, the generalizability theory enables us to estimate, at once, the average topic sample size that is required to yield sufficiently dependable estimates of the population MAPs of a given set of IR systems by means of a decision study or a D-study, which includes the method called the analysis of variance (ANOVA) in statistics.

As a matter of fact, if we were to draw a great many samples from the same population, on average, the sample standard deviations (s) would not give an unbiased estimate of σ, so the standard error $s/\sqrt{n}$ would not give an unbiased estimate of the standard deviation of the true sampling distribution, $\sigma/\sqrt{n}$. First, this is because of the square rooting and increased rounding error, and, second, $s \neq \sigma$, on average. However, a great many sample variances (s^2s) drawn from the same population will indeed give us an unbiased estimate of σ^2, i.e. the s^2's average will equal the population variance σ^2.

For the works, ANOVA included, s^2 (and not s) is used as the measure of the population spread. In ANOVA, the mean squared error (MSE), or the amount by which the estimate differs from the true value being estimated, is given by:

$$MSE = \frac{1}{df_e} \sum e_p^2,$$

where e_p is residual and df_e is the error degrees of freedom. Under the null hypothesis, the expected value of the MSE or the variance of e_p gives the model variance, i.e. $E(MSE) = \sigma^2$. For TREC 6, 7, and 8, Table 3 lists the corresponding model variances.

Table 3. MSE or the model variance σ^2 for TREC 6, 7, and 8. Aver. s^2 is the same in Table 2, repeated here for convenience.

MSE		Aver. s^2
TREC 6	0.0305	0.0305
TREC 7	0.0248	0.0248
TREC 8	0.0379	0.0379
Average	0.0311	0.0311

As shown in Table 3, the calculated MSE for each TREC corresponds to the s^2 averages given in Table 2

('Aver. s^2' in Table 3). This is the expected case, because the MSE is given by the pooled sample variance under a balanced design with (AP) samples of equal size. This means that the average topic sample size estimates given in Table 2 are actually the figures that would be obtained by conducting a D-study.

According to generalizability theory, given a set of IR systems, one can estimate the sufficient size of a topic sample as given by:

$$n \geq \frac{MSE}{\sigma^2_{err}} \quad \left[\equiv \frac{s^2_{average}}{\left(\delta / z_{\alpha/2}\right)^2} \equiv \left(\frac{s \cdot z_{\alpha/2}}{\delta} \right)^2 \right],$$

where σ^2_{err} denotes the error variance and the subscript 'average' in s^2 in square brackets means averaging over all pairs of systems available. Here, n is the sample size that is required to be $100 \times (1-\alpha)\%$ confident that the same (relative) systems ranking will be observed across the topic samples of size n. However, note that this is valid only for those system pairs that have at least a difference in population MAP of δ: for those system pairs that have a difference of less than δ, the associated ranks may still vary by chance across the systems rankings.

For instance, suppose that the decision maker wants to be 95% confident that the population MAP of a TREC 6 run is within ± 0.01 of the MAP to be observed before deciding whether a system is better for MAP than another on the population of topics, based on the systems ranking to be obtained. According to this 5% margin of error (i.e. $\alpha = 0.05$), the error variance σ^2_{err} is $(0.01/1.96)^2$, where $z_{\alpha/2} = 1.96$ (i.e. 2-sided significance). Thus, the average number of topics, which is required to estimate the true population MAPs of individual TREC 6 runs with $\delta = \pm 0.01$ sensitivity at $\alpha = 0.05$, can be estimated as given by:

$$n \geq \frac{0.0305}{(0.01/1.96)^2} \approx 1172.$$

Similarly, for a sensitivity level of $\delta = \pm 0.05$, it is approximately 47 (i.e. $0.0305/(0.05/1.96)^2$), and for $\delta = \pm 0.06$, it is approximately 33. As expected, these average topic sample size estimates agree with the figures given in Table 2. Note that, to get a 1-sided MSE-based estimate, we could use $z_{0.05}$ instead of $z_{0.025}$ for the calculation of the error variance σ^2_{err}, but this would not be appropriate for the purpose of making a decision based on the systems rankings.

Table 4 shows, for TREC 6, 7, and 8, the summary statistics of the topic sample size estimates that are yielded from the methods discussed so far, namely Inequality 2, REER, and MSE.

Recall that Voorhees and Buckley [6] concluded, based on the results of the original REER algorithm, that an absolute difference in MAP of 0.05–0.06 would be needed between 2 IR systems measured on 50 topics before concluding, with 95% confidence, that the same systems ranking can be obtained on a different set of 50 topics. As seen in Table 4, the theoretical REER approximation based on Inequality 2 with s_p and $z_{0.05} = 1.64$ ('Inq2(s_p,1.64)') yields, on average, similar estimates (≥ 0.0443), as well as the MSE-based estimation at $\delta = \pm 0.05$ ('Inq2(MSE,1.96)').

However, note that ANOVA (a MSE-based estimation) assumes that the distributions of the AP scores of individual IR systems are mutually independent of the population of topics, as also assumed by the theoretic REER approximation. Thus, the required size of a topic sample is, in fact, equal to half of the total sample size suggested, as shown in the last 2 rows 'REER/2' and 'MSE/2', which are approximately equal to the topic sample size estimates yielded from Inequality 2 with s_d and $z_{0.05} = 1.64$ ('Inq2(s_d,1.64)'). Note that the correspondence among those 3 types of estimations actually suggests that the similar estimates can also be obtained by means of Inequality 1 ('Aver. s^2') with minimal effort.

Table 4. Summary statistics of the topic sample size estimates yielded from Inequality 2 for TREC 6, 7, and 8. For each TREC, 'Run pairs' lists the number of run pairs between each of which the observed MAP difference falls within the intervals shown in columns (≥ 0 and < 1, ≥ 1 and < 2, and so on), and 'Aver. diff.' gives the average MAP difference over those run pairs; similarly, 'Inq2(s_d,1.64)' gives the average of the topic sample size estimates yielded from Inequality 2 with s_d and $z_{0.05} = 1.64$, and 'Inq2(s_p,1.64)' the same with s_p, i.e. REER. 'Inq2(MSE,1.96)' lists the MSE-based estimates with $z_{0.05} = 1.96$ at $\delta = \pm 0.01, \pm 0.02, \ldots, \pm 0.07$. The estimates listed in 'Aver. s^2' are the same in Table 2, repeated here for convenience. Grand averages over 3 TRECs are given in the row block 'Averages'. The last 2 rows list half of the estimates based on the REER and MSE.

$\lvert\bar{x} - \bar{y}\rvert$		≥ 0.0	≥ 0.01	≥ 0.02	≥ 0.03	≥ 0.04	≥ 0.05	≥ 0.06
δ		± 0.01	± 0.02	± 0.03	± 0.04	$\pm$**0.05**	± 0.06	± 0.07
TREC 6	Run pairs	54	54	44	44	30	32	148
	Aver. diff.	0.0050	0.0153	0.0248	0.0348	0.0440	0.0547	0.1215
	Inq2(s_d,1.64)	53700	281	89	56	30	24	8
(REER)	Inq2(s_p,1.64)	175634	472	156	82	50	28	8
	Inq2(MSE,1.96)	1172	292	130	73	47	33	24
	'Aver. s^2'	1175	294	131	73	47	33	24
TREC 7	Run pairs	9	4	7	11	11	6	105
	Aver. diff.	0.0043	0.0157	0.0252	0.0347	0.0441	0.0558	0.1394
	Inq2(s_d,1.64)	3478	244	99	32	24	19	5
(REER)	Inq2(s_p,1.64)	8562	282	122	53	39	24	6
	Inq2(MSE,1.96)	953	238	106	60	38	26	19
	'Aver. s^2'	953	238	106	60	38	26	19
TREC 8	Run pairs	130	128	123	110	96	85	812
	Aver. diff.	0.0052	0.0149	0.0252	0.0345	0.0448	0.0552	0.1614
	Inq2(s_d,1.64)	67434	234	69	35	23	16	5
(REER)	Inq2(s_p,1.64)	184946	626	195	104	58	40	7
	Inq2(MSE,1.96)	1456	364	162	91	58	40	30
	'Aver. s^2'	1457	364	162	91	58	40	30
Averages	Run pairs	64	62	58	55	46	41	355
	Aver. diff.	0.0048	0.0153	0.0251	**0.0347**	**0.0443**	0.0552	0.1408
	Inq2(s_d,1.64)	41537	253	86	**41**	26	20	6
(REER)	Inq2(s_p,1.64)	123047	460	158	80	**49**	31	7
	Inq2(MSE,1.96)	1194	298	133	75	**48**	33	24
	'Aver. s^2'	1195	299	133	75	48	33	24
	REER/2	61524	230	79	**40**	25	15	4
	MSE/2	597	149	66	**37**	24	17	12

As a result, for any pair of TREC runs with equal population MAPs, a MAP difference as extreme as or more extreme than 0.0347 is expected on, at most, 5% of the samples of 50 topics that could be drawn from the population. Thus, given a sample of 50 topics, an absolute difference in MAP of 0.0347 would, on average, be enough to provide 95% confidence on a declared significance between 2 TREC runs. More precisely, given a sample of 50 topics, $Pr(|\mu_X - \mu_Y| > 0) \geq 0.95$,if $|\bar{x} - \bar{y}| \geq 0.0347$, because when $\mu_X = \mu_Y$, $Pr(|\bar{x} - \bar{y}| \geq 0.0347) \leq 0.05$.

Table 5 lists, for each of the pairs of the 12 selected TREC 6 runs, the topic sample size required to provide 95% confidence on a declared significance, where the average topic sample size is 536 over the total 66 run pairs listed. As seen, for the individual pairs of TREC 6 runs, the topic sample size that is sufficient to have a confidence level of 95% varies from a size as small as 10 ('Cor6A1cls' vs. 'Brkly21' with a difference in

MAP of 0.0423) to a size as large as 722 ('gmu97au1' vs. 'pirc7Ad' with a difference in MAP of 0.0128) when the marginal MAP differences that are less than 0.01 are not considered.

Table 5. Topic sample sizes suggested by Inequality 2 at $\alpha = 0.05$ to provide 95% confidence on a declared significance. The lower triangle lists the observed MAP differences and the upper triangle lists the corresponding topic sample size estimates. The content of the cell associated with each run pair to which the 1-tailed, paired t-test gives significance at $\alpha = 0.05$ (i.e. a declared significance) is bold-faced.

Runs	MAP	Rank	1	3	5	7	9	11
city6ad	0.2164	1		**46**	70	64	**40**	**42**
uwmt6a2	0.1912	3	**0.0252**		600	488	167	130
Cor6A2qtcs	0.1809	5	0.0355	0.0103		3255	2654	324
Cor6A1cls	0.1799	7	0.0365	0.0113	0.0010		4762	330
ibms97a	0.1775	9	**0.0389**	0.0137	0.0034	0.0024		355
gmu97au1	0.1660	11	**0.0504**	0.0252	0.0149	0.0139	0.0115	
Mercure2	0.1640	13	**0.0524**	0.0272	0.0169	0.0159	0.0135	0.0020
pirc7Ad	0.1533	15	**0.0632**	0.0380	0.0277	0.0267	0.0243	0.0128
umcpa197	0.1460	17	**0.0704**	**0.0452**	**0.0349**	**0.0339**	**0.0315**	0.0200
Brkly21	0.1376	19	**0.0789**	**0.0536**	**0.0433**	**0.0423**	**0.0400**	0.0285
DCU97snt	0.1296	21	**0.0868**	**0.0616**	**0.0513**	**0.0503**	**0.0479**	0.0364
csiro97a2	0.1172	23	**0.0993**	**0.0740**	**0.0638**	**0.0627**	**0.0604**	**0.0489**
Runs	MAP	Rank	13	15	17	19	21	23
city6ad	0.2164	1	**27**	**16**	**14**	**13**	**16**	**12**
uwmt6a2	0.1912	3	56	51	**20**	**19**	**25**	**14**
Cor6A2qtcs	0.1809	5	55	100	**22**	**12**	**40**	**12**
Cor6A1cls	0.1799	7	82	103	**26**	**10**	**42**	**15**
ibms97a	0.1775	9	115	121	**28**	**14**	**40**	**15**
gmu97au1	0.1660	11	14284	722	164	59	98	**40**
Mercure2	0.1640	13		627	**48**	**37**	75	**13**
pirc7Ad	0.1533	15	0.0107		1305	290	101	66
umcpa197	0.1460	17	**0.0180**	0.0072		332	309	**36**
Brkly21	0.1376	19	**0.0264**	0.0157	0.0084		1574	144
DCU97snt	0.1296	21	0.0344	0.0236	0.0164	0.0080		564
csiro97a2	0.1172	23	**0.0468**	0.0361	**0.0289**	0.0204	0.0124	

4. Conclusion

In this article, the second fundamental theorem of probability, the CLT, is exploited for the empirical estimation of the sufficient size of a topic sample. To this extent, this article can be considered as the detailed and formal explanation of the research methodology that should be followed in the design of IR experiments when the methods of statistical inference are used to give significance to the results of an IR system evaluation that follows the Cranfield paradigm.

The results of the statistical analyses performed show that, if the null hypothesis H_0 of equal population MAPs is rejected for 2 IR systems based on a sample of 50 topics, an absolute difference in MAP of 0.03 or more would actually be enough to ensure that the chance of rejecting H_0 when it is true is at or below the nominal level of 5%. Previous empirical research consistently singled out a MAP difference that is not less than 0.05, simply because it was assumed that the AP distributions of IR systems are independent of each other on the population of topics, whereas it depends on the IR systems under consideration. On average, a sample of 25 topics or less can indeed be enough to provide 95% confidence on a declared significance for 2 IR systems

having an observed MAP difference of 0.05 or more, while a sample of 50 topics or more will be necessary to maintain the same level of confidence for those IR systems that have an observed MAP difference of 0.03 or less.

Holding those average figures on one hand, on the other hand, the statistical analyses performed also revealed that the sufficient size of a topic sample greatly varies from system pair to system pair in practice. In TREC 6, for example, the topic sample size required to provide 95% confidence on a declared significance ranges from a size as small as 10 to a size as large as 722, where the corresponding MAP differences range from 0.1 down to 0.01, respectively. Thus, when system pairs are considered separately, it can be said that a standard TREC test collection with 50 topics could succeed to provide the necessary but probably not the sufficient empirical basis to detect important population effects among IR systems. It will therefore be better if every pair of IR systems is considered as a separate case with respect to the design of IR experiments, rather than relying on average figures as a rule of thumb.

References

[1] C.W. Cleverdon, "The significance of the Cranfield tests on index languages", Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (reprint), pp. 3–12, 1991.

[2] W.G. Cochran, Sampling Techniques, New York, Wiley, 1977.

[3] R.V. Hogg, A.T. Craig, J.W. McKean, Introduction to Mathematical Statistics, New York, Prentice Hall, 2004.

[4] K. Sparck Jones, "Automatic indexing", Journal of Documentation, Vol. 30, pp. 393–432, 1974.

[5] C. Buckley, E.M. Voorhees, "Evaluating evaluation measure stability", Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 33–40, 2000.

[6] E.M. Voorhees, C. Buckley, "The effect of topic set size on retrieval experiment error", Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 316–323, 2002.

[7] W. Webber, A. Moffat, J. Zobel, "Statistical power in retrieval experimentation", Proceeding of the 17th ACM conference on Information and Knowledge Management, pp. 571–580, 2008.

[8] E.M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness", Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 315–323, 1998.

[9] E.M. Voorhees, "Topic set size redux", Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 806–807, 2009.

[10] R.A. Fisher, The Design of Experiments, Edinburgh, Oliver & Boyd, 1935.

[11] W. Lin, A. Hauptmann, "Revisiting the effect of topic set size on retrieval error", Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 637–638, 2005.

[12] D. Bodoff, P. Li, "Test theory for assessing IR test collections", Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 367–374, 2007.