# Confidence Intervals for Information Retrieval Evaluation

*Laurence A. F. Park*
School of Computing and Mathematics
University of Western Sydney, Australia
*lapark@scm.uws.edu.au*

**Abstract**
*Information retrieval results are currently limited to the publication in which they exist. Significance tests are used to remove the dependence of the evaluation on the query sample, but the findings cannot be transferred to other systems not involved in the test. Confidence intervals for the population parameters provide query independent results and give insight to how each system is expected to behave when queried. Confidence intervals also allow the reader to compare results across articles because they provide the possible location of a systems population parameter. Unfortunately, we can only construct confidence intervals of population parameters if we have knowledge of the evaluation score distribution for each system. In this article, we investigate the distribution of Average Precision of a set of systems and examine if we can construct confidence intervals for the population mean Average Precision with a given level of confidence. We found that by standardising the scores, the system score distribution and system score sample mean distribution was approximately Normal for all systems, allowing us to construct accurate confidence intervals for the population mean Average Precision.*

**Keywords** Information Retrieval, Evaluation

## 1 Introduction

When publishing information retrieval system evaluation results, the mean score from a sample set of queries is reported. These results are usually presented with the confidence in hypothesis test results when compared with a baseline system. Reporting the sample mean allows the reader to compare the presented set of systems for the given set of queries, while the hypothesis tests indicate how well the results generalise to a new sample of queries.

Unfortunately, there is no method for comparing systems across publications. We are able to compare the sample mean scores, but by doing so we have no indication of how the systems will perform when given a new sample of queries. Results from hypothesis test report the confidence in the test, and therefore the tests information cannot be used to compare systems across publications. The reader's only option is to obtain the set of systems in each publication and run experiments to identify if one is more accurate than the other.

By having knowledge of a population parameter, such as the population mean evaluation score for a system, we would be able to compare systems independent of the sample set of queries used. We are unable to compute an exact value for the population mean using a sample set of queries, but we are able to construct a confidence interval, giving a range in which the population mean evaluation score is most likely to exist.

To compute accurate confidence intervals for a population parameter from samples, we must have knowledge of the distribution of the associated sample statistic. In this article, we investigate the distribution of Average Precision and the sample mean Average Precision to compute accurate confidence intervals for the population mean Average Precision.

We make the following important contributions:

- An investigation into how we can report the confidence intervals for the population mean Average Precision (Section 4 and 5).
- A description on how results should be reported to allow others to reuse the results (Section 6).

The article will proceed as follows: Section 2 provides a brief overview of Information Retrieval evaluation, Section 3 discusses the portability of published Information Retrieval results, Section 4 examines the distribution of Average Precision results and identifies if we are able to construct accurate confidence intervals. Section 5 examines the effect of standardisation of the distribution of Average Precision results. Finally, Section 6 presents further details of the confidence interval we have found.

## 2 System evaluation

First let us define the retrieval system. A retrieval system is a function $S(q, D)$ on query $q$ and document set $D$, where $S : q \times D \rightarrow \mathbb{R}^N$. The output of the function $S$ is a vector $\vec{r}_{q,D} = \{r_{q,d_1}, r_{q,d_2}, \ldots, r_{q,d_N}\}$ containing a weighted list, where each weight $r_{q,d_i}$ is associated to the relevance of document $i$ in $D$ to query $q$.

An evaluation measure is a function $m_{q,D} = E(\vec{r}_{q,D}, \vec{\rho}_{q,D})$ on the weighted document list $\vec{r}_{q,D}$ and the set of true relevance judgements $\vec{\rho}_{q,D}$, where $E : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$. The output of $E$ is a scalar value

**Proceedings of the 15th Australasian Document Computing Symposium, Melbourne, Australia, 10 December 2010.**

which reflects the accuracy of system $S$ on document set $D$ using query $q$.

To truly test the accuracy of a system on a document collection, we would obtain all of the queries that will be used, along with their probability of use, and compute the expected system accuracy using

$$\mathbb{E}[s_D] = \sum_{q \in \Phi} E(\vec{r}_{q,D}, \vec{\rho}_{q,D}) P(q)$$

where $P(q)$ is the probability of issuing query $q$ and $\Phi$ is the population of queries.

Two problems exist with this form of evaluation. First, the population of queries $\Phi$ depends on the future use of the system. We could obtain an estimate of $\Phi$ by releasing the system and recording all of the queries that are used, but there is no way of knowing how good an estimate this is. Second, for each query we need a set of relevance judgements for document set $D$. For one query, this requires manually judging the relevance of all documents in $D$. If $D$ contains one million documents, we must perform one million relevance judgements. For $k$ queries, we must perform $k$ million judgements.

To overcome the first problem, the information retrieval community has resorted to using a sample of queries and treating each query as being equally likely. This changes the expectation equation to simply computing the mean of a sample:

$$\overline{m}_D = \sum_{q \in Q} E(\vec{r}_{q,D}, \vec{\rho}_{q,D}) \frac{1}{k}$$

where $Q \subset \Phi$, $k$ is the cardinality of set $Q$, and $\overline{m}_D$ is the sample mean system score over the query sample set $Q$. The sample mean is used as an estimate of the population mean (expected value), but estimates of how well this is approximated are not provided in experimental results.

To overcome the second problem, methods such as pooling [3] can be used to reduce the load of this task, but significant effort must still be placed into this process.

By themselves, the sample mean evaluation scores are limited in their use. The sample mean scores are used in most retrieval experiments to compare against the sample mean retrieval scores of another system, where both systems are evaluated using the same sample.

To remove the dependence of the evaluation on the query sample, a paired hypothesis test (e.g. the Wilcoxon signed rank test) for an increase in evaluation score can be performed for a pair of systems. The result from the test is the level of confidence of the first system providing a greater score than the second for a randomly sampled query.

## 3 Portability of results

We showed in the previous section that we are able to compare two retrieval systems using a paired significance test. To conduct the test, we require the evaluation score for each system for a specific set of queries. Therefore, if we have access to both systems $S_x$ and $S_y$, and we have a document set $D$, a random sample of queries $Q$ and the associated relevance judgements, we simply generate the system score using a suitable evaluation metric $E$ and compare the paired evaluation scores using a significance test.

If a reader obtains two publications that have developed new systems $S_x$ and $S_y$ respectively, the reader is unable to determine from the published results in both articles if there is any statistically significant difference in results between systems $S_x$ and $S_y$. The reader should be able to compare the sample means of each system from each article as an estimate of the expected performance of each system, but the reader would have no knowledge of the accuracy of the estimation. Paired significance tests would be provided in each article, but the paired test results only apply to the systems involved in the test and give no indication of how the system compares with others not involved in the test.

At the moment, the only way to compare two systems that appear in separate publications is to obtain the systems and run our own experiments. This implies that the current method of reporting information retrieval results limits the evaluation to the publication. We are unable to compare retrieval evaluations across articles and therefore our results are not portable.

To provide portable results, all retrieval experiments should provide details of system population parameters. Population parameters provide details on how the evaluated system behaves independent of the query sample used and can also provide us with information such as the expected evaluation score for the system.

Since system population parameters are independent of the query sample, we are able to compare the values of multiple systems across different publications, making the results portable.

If we obtain a sample from a given population, we are not able to compute the exact value of population parameters, but we can compute a confidence interval for a population parameter using statistical methods. Therefore, if we have a set of evaluation scores for a given system obtained from a sample set of queries, we are able to compute a confidence interval for a certain population parameter.

For each confidence interval, we need an associated confidence level, where the confidence level is related to the probability of a Type I error occurring (the probability of the population parameter not being in the interval). For a confidence interval to be useful, the probability of a Type I error should be low.

To accurately compute the probability of a Type I error for a given confidence interval, we need to know the distribution function associated to the sample data. Therefore to compute the confidence level of a confidence interval for a given system, we need to know the distribution function of the evaluation score distribution. To the best of our knowledge, there has been no study into the distribution of retrieval system evaluation scores.

In the following sections we will investigate the distribution of Average Precision over a set of systems and identify how we can use the distribution to construct a confidence interval for the population mean Average Precision with an associated accurate measure of Type I error.

A system's population mean evaluation score is the expected score for a randomly sampled query. This parameter is of interest because it provides us with a measure of how well the system will perform when provided with an unknown query. There has been much research into computing the confidence interval of the population mean for given distributions, therefore we will use the knowledge from the prior research and identify how well it applies to a set of system distributions.

To compute the confidence interval for a system population mean Average Precision (AP), we must

1. identify the distribution of the sample mean AP,
2. compute an estimate of the parameters of the sample mean AP distribution given the sample,
3. finally, identify the quantiles of the distribution that contain the desired level of confidence.

## 4 Average Precision Distribution

To test the validity of confidence interval experiments, we require knowledge of the population statistics of the system score distributions. A system score distribution is the probability of obtaining a particular score from a randomly sampled query for the given retrieval system on a given document set. System score distributions have not been computed or approximated for any retrieval system (to the best of our knowledge). Therefore we will approximate a set of system score distributions using the scores from a large sample of queries.

In this article, we have used the system scores from the TREC 2004 Robust track. The TREC Robust track contains 249 queries and results from 110 retrieval systems on a document collection containing $528,155$ documents from TREC disks 4 and 5 (excluding the Congressional Record document set).

We will use the following notation:

- AP is the Average Precision from a sample query for a given system,
- $\overline{\text{AP}}$ is the sample mean Average Precision for a given system from a sample of $n$ queries (usually known as mean Average Precision),
- $\mu_{\text{AP}}$ is the population mean Average Precision for a given system,
- $s_{\text{AP}}$ is the sample standard deviation Average Precision for given system from a sample of $n$ queries,
- $\sigma_{\text{AP}}$ is the population standard deviation of Average Precision for a given system,

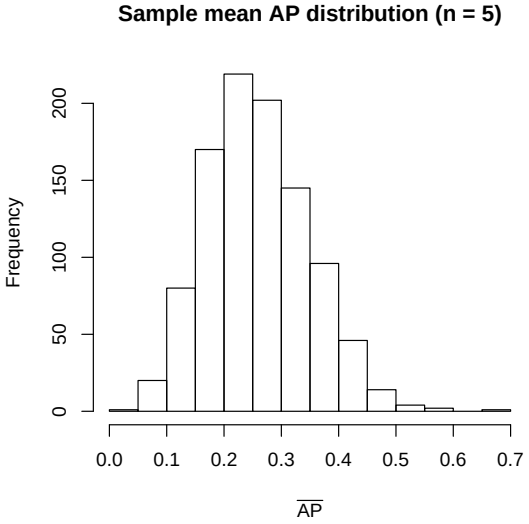


Figure 1: Distribution of a randomly sampled system's sample mean ($\overline{\text{AP}}$) using $n = 5$.

- $\sigma_{\overline{\text{AP}}}$ is the population standard deviation of the sample mean Average Precision for a given system.

Using the TREC Robust data, we are able to estimate the population parameters using the set of 249 queries and the sample statistics using a smaller subset of the queries. For example, $\mu_{\text{AP}}$ is computed for a given system by computing the mean across all 249 queries, while $\overline{\text{AP}}$ is computed using a small subset of the queries (such as $n = 10$).

### 4.1 Confidence when $\sigma_{\text{AP}}$ is known

In this section we will examine the accuracy of a confidence interval under the assumption that AP follows a Normal distribution and $\sigma_{\text{AP}}$ is known for each system.

The Central Limit Theorem [2] tells us that given a Normally distributed random variable $x$ with mean $\mu$ and standard deviation $\sigma$, its sample mean $\overline{x}$ is also Normally distributed with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$, where $n$ is the number of samples taken.

The Central Limit Theorem also tells us that if $x$ is not Normally distributed, but our sample size, $n$ is large ($n > 30$), then the sample mean is approximately Normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$.

A histogram of a typical system's $\overline{\text{AP}}$ is shown in Figure 1. It shows that the sample mean is approximately Normal. This is also the case for all other systems. Therefore, to begin, we will assume that a system's $\overline{\text{AP}}$ follows a Normal distribution, where each system distribution is characterised by its mean $\mu_{\text{AP}}$ and standard deviation $\sigma_{\text{AP}}$.

We will also assume that we know each systems standard deviation ($\sigma_{\text{AP}}$). This is not a useful assumption in practice, but it will allow us to investigate if our assumption of Normality is valid.

Given that $\overline{\text{AP}}$ is Normal for each system, we can compute the confidence interval of $\mu_{\text{AP}}$ using:

$$\mu_{\text{AP}} \in \overline{\text{AP}} \pm Z_{\alpha/2}\sigma_{\text{AP}}/\sqrt{n} \quad (1)$$

where $\alpha \in [0, 1]$ is the probability of a Type I error, the level of confidence is $100(1-\alpha)\%$, and $Z_{\alpha/2}$ is the $\alpha/2$ quantile of the Standard Normal distribution (meaning that $100(1-\alpha)\%$ of the Standard Normal distribution lies between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$).

Our first experiments examines the Type I error ($\alpha$) of the confidence interval. Using a set of 110 system scores, we compute an estimate of $\mu_{\text{AP}}$ and $\sigma_{\text{AP}}$ using the AP results from all 249 queries. By taking a random sample of $n = 5$ AP scores for a particular system, we are able to compute the confidence interval of $\mu_{\text{AP}}$ and compare it to the our computed value of $\mu_{\text{AP}}$. If $\mu_{\text{AP}}$ does not lie within the confidence interval, a Type I error has occurred. The value of $\alpha$ provided in the confidence interval calculations is the expected Type I error. Therefore, repeated experiments should show the Type I error of the confidence interval to be equal to $\alpha$. For our experiment, we computed 1000 confidence intervals for each system, from random samples of $n = 5$ AP scores. The results are presented in Table 1

Table 1: The actual Type I error produced when computing $\mu_{\text{AP}}$ confidence intervals using knowledge of $\sigma_{\text{AP}}$, given $\alpha$. The mean, standard deviation and maximum across all systems are computed from 1000 confidence intervals using $n = 5$ for each system.

| $\alpha$ | Type I error | | |
|---|---|---|---|
| | Mean | SD | Max |
| 0.050 | 0.040 | 0.005 | 0.051 |
| 0.100 | 0.088 | 0.009 | 0.109 |
| 0.150 | 0.139 | 0.012 | 0.161 |
| 0.200 | 0.191 | 0.014 | 0.208 |
| 0.250 | 0.242 | 0.014 | 0.269 |
| 0.300 | 0.295 | 0.013 | 0.320 |
| 0.350 | 0.348 | 0.011 | 0.376 |
| 0.400 | 0.400 | 0.010 | 0.424 |
| 0.450 | 0.452 | 0.010 | 0.482 |
| 0.500 | 0.502 | 0.010 | 0.533 |

If the system sample mean distributions are Normal, we would expect to see that the Type I error from all systems be close to the given $\alpha$. The results show that the Type I error across all systems is close to the value of $\alpha$ implying that the confidence interval being used is correct. Similar results are obtained when using other values of $n$. These results imply that our assumption that $\overline{\text{AP}}$ is Normal is valid.

## 4.2 Confidence when $\sigma_{\text{AP}}$ is unknown

In the previous section we assumed that $\sigma_{\text{AP}}$ is known, which would not be the case when estimating a confidence interval, but it allowed us to examine the assumption that $\overline{\text{AP}}$ followed an approximate Normal distribution.

In this section, we assume that $\sigma_{\text{AP}}$ is unknown and therefore we must approximate its value with our sample standard deviation $s_{\text{AP}}$. If AP follows a Normal distribution, then Cochran's theorem [1] provides us with:

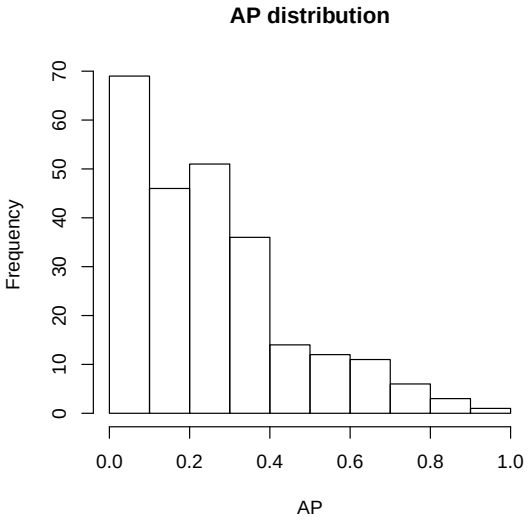


Figure 2: Distribution of a randomly sampled system's scores (AP).

$$\frac{(n-1)s^2_{\text{AP}}}{\sigma^2_{\text{AP}}} \sim \chi^2_{n-1} \quad (2)$$

where $\sim$ infers equality of distributions and $\chi^2_{n-1}$ is the Chi-squared distribution with $n-1$ degrees of freedom.

Figure 2 shows a typical system AP distribution, which does not look Normal, which may infer that the relationship in Cochran's theorem is not valid. The Q-Q plot in Figure 3 shows that the relationship in Cochran's theorem is valid except at the higher end of the scale, implying that the $\chi^2$ distribution has a longer tail than the variance ratio ($(n-1)s^2_{\text{AP}}/\sigma^2_{\text{AP}}$). This implies that score samples with high standard deviation will provide an under estimate of the confidence interval.

By estimating $\sigma_{\text{AP}}$ with $s_{\text{AP}}$ using the relationship in equation 2, we arrive at the confidence interval relationship:

$$\mu_{\text{AP}} \in \overline{\text{AP}} \pm t_{\alpha/2,n-1}s_{\text{AP}}/\sqrt{n} \quad (3)$$

where $t_{\alpha/2,n-1}$ is the $\alpha/2$ quantile of the Student's $t$ distribution with $n-1$ degrees of freedom (meaning that $100(1-\alpha)\%$ of the $t$ distribution lies between $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$).

Table 2 shows the results from computing 1000 confidence intervals for each system from samples of $n = 5$ scores, using equation 3. Note that if the system scores were Normally distributed, the computed Type I error would be similar to the given $\alpha$. We can see that The mean Type I error is greater than $\alpha$ implying that we are under estimating the confidence interval width. The column providing the maximum Type I error shows a large underestimate of the confidence interval. This can be explained from our observation of the Q-Q plot

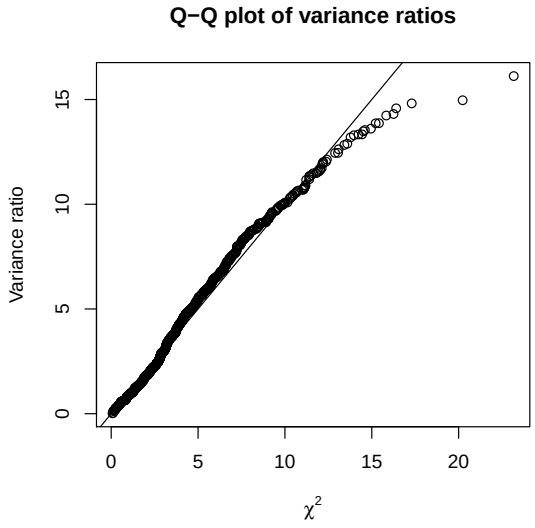


Figure 3: The Q-Q plot of the $\chi^2_{n-1}$ distribution against the $(n-1)s^2_{\text{AP}}/\sigma^2_{\text{AP}}$ distribution, for $n = 5$.

in Figure 3, showing that the samples that had larger variance do not follow the $\chi^2_{n-1}$ distribution.

Table 2: The actual Type I error produced when computing $\mu_{\text{AP}}$ confidence intervals using $s_{\text{AP}}$, given $\alpha$. The mean, standard deviation and maximum across all systems are computed from 1000 confidence intervals using $n = 5$ for each system.

| $\alpha$ | Type I error | | |
|---|---|---|---|
| | Mean | SD | Max |
| 0.05 | 0.082 | 0.027 | 0.255 |
| 0.10 | 0.133 | 0.027 | 0.299 |
| 0.15 | 0.179 | 0.026 | 0.340 |
| 0.20 | 0.224 | 0.024 | 0.377 |
| 0.25 | 0.269 | 0.021 | 0.407 |
| 0.30 | 0.315 | 0.020 | 0.440 |
| 0.35 | 0.362 | 0.018 | 0.480 |
| 0.40 | 0.410 | 0.018 | 0.520 |
| 0.45 | 0.458 | 0.018 | 0.559 |
| 0.50 | 0.504 | 0.018 | 0.602 |

We now have the problem that we are unable to obtain a good estimate of the score population standard deviation $\sigma_{\text{AP}}$ and hence unable to obtain an accurate confidence interval for $\mu_{\text{AP}}$ from a sample of scores. To proceed, we must either obtain the distribution of $(n-1)s^2_{\text{AP}}/\sigma^2_{\text{AP}}$, or find a mapping that provides us with Normally distributed AP. In the next section, we will examine the latter using score standardisation.

# 5 Standardised AP

Score standardisation was introduced as a method of allowing cross collection comparison of system scores [4]. In this section, we will examine the effect of standardisation on the distribution of AP and its effect on confidence interval estimations.

Standardised AP is defined as:

$$\text{sAP}_q = \frac{\text{AP}_q - \overline{\text{AP}}_q}{s_{\text{AP},q}}$$

where $\text{sAP}_q$ is the standardised AP for a given system on query $q$, $\text{AP}_q$ is the Average Precision for the given system on query $q$, $\overline{\text{AP}}_q$ is the mean AP across a set of systems for query $q$, and $s_{\text{AP},q}$ is the standard deviation across a set of systems for query $q$. From this definition, we can see that standardisation is highly dependent on the set of systems (from which $\overline{\text{AP}}_q$ and $s_{\text{AP},q}$ are computed). Therefore, we will begin the investigation using all systems to perform the standardisation and finish by examining the effect of using a small sample to perform standardisation.

We will use the following notation:

- sAP is the standardised Average Precision from a sample query for a given system,
- $\overline{\text{sAP}}$ is the sample mean standardised Average Precision for a given system from a sample of $n$ queries,
- $\mu_{\text{sAP}}$ is the population mean standardised Average Precision for a given system,
- $s_{\text{sAP}}$ is the sample standard deviation standardised Average Precision for given system from a sample of $n$ queries,
- $\sigma_{\text{sAP}}$ is the population standard deviation standardsed Average Precision for a given system,
- $\sigma_{\overline{\text{sAP}}}$ is the population standard deviation of the sample mean standardised Average Precision for a given system.

where $\mu_{\text{sAP}}$ and $\sigma_{\text{sAP}}$ are estimated using all 249 queries.

## 5.1 Standardisation using all systems

In this section we will use all 110 systems to compute the mean and standard deviation of each query to perform standardisation. Note that when performing retrieval experiments, it would be unlikely to have evaluated 110 systems on the set of queries being evaluated. Therefore, this section is similar to a 'best case' analysis. We also present the confidence intervals for when $\sigma_{\text{sAP}}$ is known and unknown to identify where any problems in our assumptions lie.

### 5.1.1 Confidence when $\sigma_{\text{sAP}}$ is known

By performing the standardisation, we obtain a sAP score for each query. To establish the confidence interval for $\mu_{\text{sAP}}$, we must deduce the distribution of $\overline{\text{sAP}}$. A histogram of the distribution of a system's $\overline{\text{sAP}}$ is shown in Figure 4. We can see that the particular system sample mean sAP is approximately Normal. If we

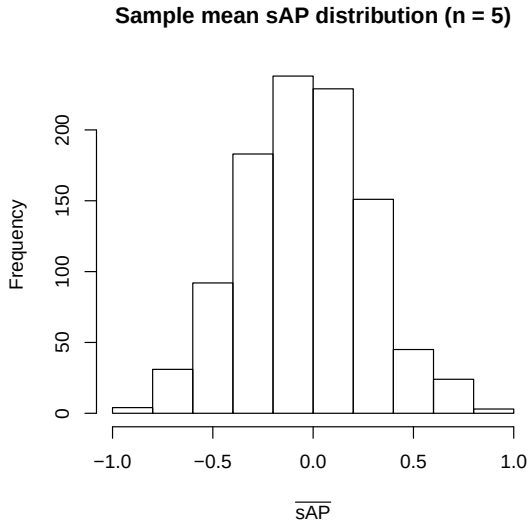


Figure 4: Distribution of a randomly sampled system's sample mean ($\overline{\text{sAP}}$) using $n = 5$.

examine the $\overline{\text{AP}}$ distribution in Figure 1, we find that the $\overline{\text{sAP}}$ distribution is less skewed giving it a more Normal appearance. This Normality implies that we should obtain accurate confidence intervals when the system population standard deviation $\sigma_{\text{sAP}}$ is known.

To compute the accuracy of the confidence interval estimates when $\sigma_{\text{sAP}}$ is known, we used equation 1 and replaced AP with sAP. Samples of size $n = 5$ were used to compute the confidence interval and compared to $\mu_{\text{sAP}}$. If $\mu_{\text{sAP}}$ was not in the confidence interval, a Type I error occurred. This was repeated 1000 times for each system. The probability of a Type I error is listed in Table 3. Table 3 reports the mean, standard deviation and maximum probability of a Type I error across all systems. The table shows mean and maximum values similar to the associated values of $\alpha$, and small standard deviation. This implies that the confidence intervals produced are accurate.

#### 5.1.2 Confidence when $\sigma_{\text{sAP}}$ is unknown

We have found that the Normal distribution is a good approximation for the distribution of $\overline{\text{sAP}}$. In this section we will examine if we can approximate $\sigma_{\text{sAP}}$ using $s_{\text{sAP}}$ and Cochran's theorem (equation 2).

Cochran's theorem is valid under the assumption that the data follows a Normal distribution. The histogram of a sample system's sAP in Figure 5 shows that sAP is approximately Normal. To examine if this approximation is close, we have also examined the Q-Q plot of the variance ratio on the left hand side of equation 2 compared to the $\chi^2$ distribution on the right hand side of equation 2. The plot (given in Figure 6) shows that the two distributions are approximately equal, suggesting that we are able to use $s_{\text{sAP}}$ to approximate $\sigma_{\text{sAP}}$.

The confidence interval is computed using equation 3, where we replace all occurrences of AP with sAP. We investigated the accuracy of the confidence interval

Table 3: The actual Type I error produced when computing $\mu_{\text{sAP}}$ confidence intervals using $\sigma_{\text{sAP}}$, given $\alpha$. The mean, standard deviation and maximum across all systems are computed from 1000 confidence intervals using $n = 5$ for each system.

| $\alpha$ | Type I error | | |
|---|---|---|---|
| | Mean | SD | Max |
| 0.050 | 0.046 | 0.006 | 0.062 |
| 0.100 | 0.094 | 0.009 | 0.118 |
| 0.150 | 0.142 | 0.012 | 0.175 |
| 0.200 | 0.192 | 0.014 | 0.221 |
| 0.250 | 0.243 | 0.015 | 0.273 |
| 0.300 | 0.292 | 0.016 | 0.324 |
| 0.350 | 0.342 | 0.016 | 0.376 |
| 0.400 | 0.392 | 0.016 | 0.424 |
| 0.450 | 0.443 | 0.017 | 0.474 |
| 0.500 | 0.493 | 0.017 | 0.533 |

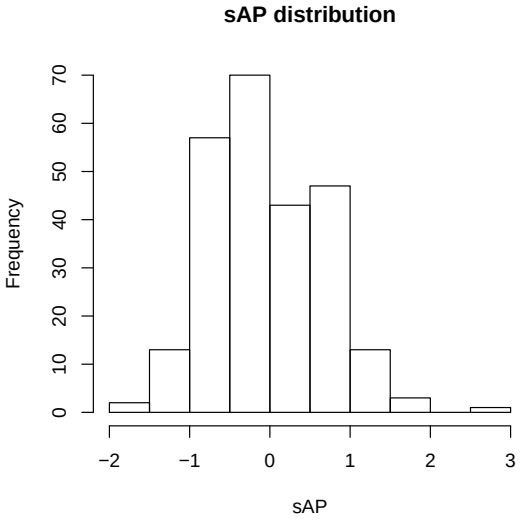


Figure 5: Distribution of a randomly sampled system's scores (sAP).

by computing the confidence interval for 1000 samples of $n = 5$ queries for each system for varying levels of $\alpha$. Statistics of the Type I error are reported in Table 4. We can see that the expected Type I error (mean) is close to the given $\alpha$, showing that the confidence interval is accurate.

### 5.2 Standardisation using a few systems

We mentioned in the previous section that it is unlikely that we would have the results from 110 systems to perform standardisation. Therefore in this section, we will examine the effect of using a random sample of five systems to perform standardisation.

To test the accuracy of our confidence intervals, we ran the same Type I error experiment from Section 5.1.2 except we used only five randomly sampled systems

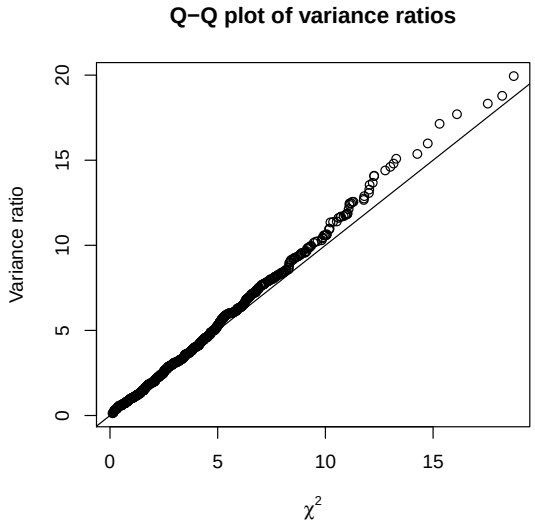


Figure 6: The Q-Q plot of the $\chi^2_{n-1}$ distribution against the $(n-1)s^2_{\text{sAP}}/\sigma^2_{\text{sAP}}$ distribution, for $n = 5$.

Table 4: The actual Type I error produced when computing $\mu_{\text{sAP}}$ confidence intervals using $s_{\text{sAP}}$, given $\alpha$. The mean, standard deviation and maximum across all systems are computed from 1000 confidence intervals using $n = 5$ for each system.

| $\alpha$ | Type I error | | |
|---|---|---|---|
| | Mean | SD | Max |
| 0.050 | 0.050 | 0.010 | 0.090 |
| 0.100 | 0.097 | 0.011 | 0.143 |
| 0.150 | 0.146 | 0.012 | 0.195 |
| 0.200 | 0.195 | 0.014 | 0.248 |
| 0.250 | 0.245 | 0.016 | 0.294 |
| 0.300 | 0.295 | 0.017 | 0.334 |
| 0.350 | 0.346 | 0.018 | 0.383 |
| 0.400 | 0.397 | 0.019 | 0.441 |
| 0.450 | 0.448 | 0.020 | 0.508 |
| 0.500 | 0.499 | 0.020 | 0.559 |

for standardisation. The results from the experiment are shown in Table 5. We can see that the expected (mean) Type I error follows $\alpha$ closely. In comparison to Table 4, we can see that the difference between $\alpha$ and the expected Type I error has increased. We can also see that the variance has increased. Therefore, reducing the number of standardisation systems has slightly decreased the accuracy of the confidence intervals, but they are more accurate than when using AP.

Note that the population mean and standard deviation are dependent on the standardising systems chosen, therefore, we cannot compare system confidence intervals when the systems have used different standardisation systems.

Table 5: The actual Type I error produced when computing $\mu_{\text{sAP}}$ confidence intervals using $s_{\text{sAP}}$ and five randomly sampled standardisation systems, given $\alpha$. The mean, standard deviation and maximum across all systems are computed from 1000 confidence intervals using $n = 5$ for each system.

| $\alpha$ | Type I error | | |
|---|---|---|---|
| | Mean | SD | Max |
| 0.050 | 0.062 | 0.017 | 0.157 |
| 0.100 | 0.114 | 0.021 | 0.212 |
| 0.150 | 0.166 | 0.024 | 0.270 |
| 0.200 | 0.217 | 0.025 | 0.312 |
| 0.250 | 0.270 | 0.026 | 0.358 |
| 0.300 | 0.324 | 0.025 | 0.408 |
| 0.350 | 0.379 | 0.025 | 0.468 |
| 0.400 | 0.435 | 0.025 | 0.520 |
| 0.450 | 0.491 | 0.026 | 0.579 |
| 0.500 | 0.543 | 0.025 | 0.634 |

Table 6: The change in Type I error as $n$ increases, where $\alpha = 0.05$ and $\sigma_{\text{AP}}$ and $\sigma_{\text{sAP}}$ are unknown.

| $n$ | 2 | 5 | 10 | 20 | 50 |
|---|---|---|---|---|---|
| AP error | 0.081 | 0.083 | 0.066 | 0.053 | 0.029 |
| sAP error | 0.050 | 0.062 | 0.068 | 0.068 | 0.054 |

## 6 Examination of Confidence Intervals

Confidence intervals are useful for identifying the likely region in which the system population mean exists. As the interval grows, the utility decreases. E.g. we could provide a $100\%$ confidence interval for $\mu_{\text{AP}}$ as $[0, 1]$. This is accurate, but does not provide us with any information since the confidence interval covers the domain of AP. In this section, we will examine the confidence intervals that were computed in the previous sections.

For the previous experiments, we have used $n = 5$ queries. We now examine the accuracy of our confidence intervals as the number of queries $n$ increases (when $\sigma_{\text{AP}}$ and $\sigma_{\text{sAP}}$ are unknown and using five standard systems). We can see in Table 6 that the Type I error for $\mu_{\text{sAP}}$ is stable, while the Type I error for $\mu_{\text{AP}}$ reduces as $n$ increases. This can be explained by the variance ratio ($(n-1)s^2_{\text{sAP}}/\sigma^2_{\text{sAP}}$) following a $\chi^2_{n-1}$ distribution. The $t$ distribution, used to compute the confidence interval, is constructed by combining the uncertainty in $\sigma_{\text{sAP}}$ given by the $\chi^2_{n-1}$ distribution with the Standard Normal distribution in equation 1. Since the variance ratio of sAP approximately follows a $\chi^2_{n-1}$ distribution, the $t$ distribution compensates for the change in $n$. The variance ratio of AP does not follow a $\chi^2_{n-1}$ distribution, therefore the $t$ distribution poorly compensates for $n$.

The confidence interval equation (shown in Equation 3) is centered at the sample mean $\overline{\text{sAP}}$ and its width is dependent on the sample standard deviation $s_{\text{AP}}$, the error rate $\alpha$ and the number of samples $n$. The sam-

Table 7: The change confidence interval (CI) width for sAP as $n$ increases, where $\alpha = 0.05$ and $\sigma_{\text{sAP}}$ is unknown.

| $n$ | 2 | 5 | 10 | 20 | 50 |
|---|---|---|---|---|---|
| CI width | 20.217 | 3.456 | 2.172 | 1.518 | 0.981 |

ple mean and standard deviation of the system score are under our control in an experimental environment, since they are the responses we are examining. In all information retrieval experiments, we have direct control over $n$, the number of queries used in the retrieval experiment, and $\alpha$.

By increasing $\alpha$ we decrease the confidence interval, but we also decrease the confidence of the confidence interval. By increasing $n$, we decrease the confidence interval, but increasing $n$ involves using a larger query set, which (if not already available) involves building the relevance judgements for the new queries. If queries are available, they should be used to increase $n$ and obtain a narrower confidence interval that will be more useful for identifying the location of $\mu_{\text{sAP}}$.

The standard deviation of the set of all sAP scores across all 110 systems, each using 1000 different randomly sampled sets of five standardisation systems is 2.436. Therefore, Table 7 shows that we need to use $n = 10$ queries to get an expected confidence interval width that is less than the standard deviation of the samples sAP. This is not a benchmark, but simply an indicator to compare the size the confidence intervals relative to the data.

Table 8: Type I error for the $\mu_{\text{sAP}}$ confidence interval (with unknown $\sigma_{\text{sAP}}$) on the 40 systems from TREC-3, using 1000 samples of $n = 5$ queries for each system and five standardisation systems.

| $\alpha$ | Type I error | | |
|---|---|---|---|
| | Mean | SD | Max |
| 0.050 | 0.055 | 0.021 | 0.142 |
| 0.100 | 0.104 | 0.024 | 0.204 |
| 0.150 | 0.153 | 0.025 | 0.249 |
| 0.200 | 0.203 | 0.026 | 0.293 |
| 0.250 | 0.253 | 0.024 | 0.340 |
| 0.300 | 0.303 | 0.026 | 0.379 |
| 0.350 | 0.355 | 0.027 | 0.421 |
| 0.400 | 0.409 | 0.028 | 0.471 |
| 0.450 | 0.462 | 0.028 | 0.525 |
| 0.500 | 0.514 | 0.027 | 0.572 |

To test the generalisation of our results, we examined the accuracy of the confidence interval method from Section 5.2 on results from TREC-3. The results in Table 8 show an expected Type I error close to the value of $\alpha$, with small standard deviation. This implies that this method of computing confidence intervals does generalise.

To report results so that others are able to compare new systems, we need to report the sample mean and sample standard deviation of Average Precision, and the number of queries used. We also need to report which systems were used to perform standardisation. Note that these systems must be freely available systems. If others do not have access to the set of standardisation systems, the confidence intervals cannot be compared. Once these items are reported, others can compute comparative confidence intervals without access to our system, queries or relevance judgements.

## 7 Conclusion

Current forms of information Retrieval report a sample mean and the confidence obtained using paired hypothesis tests. These values provide the reader with knowledge of which system is more accurate from those taking part in the experiment. Unfortunately, these values do not provide the reader with any means of comparing systems found published in different articles. We can use the system's sample mean as an estimate of the system's population mean (expected value), but the reader has no knowledge of the accuracy of this estimate.

To compare systems across publications, we would need some indication of the systems population parameters. From sample statistics, we are able to compute a confidence interval of the population mean given a certain level of confidence, as long as the sample follows a known distribution function.

In this article, we investigated the distribution of Average Precision for a set of systems and examined if we could construct accurate confidence intervals of the population mean Average Precision from a system's sample statistics.

We found that accurate confidence interval could be constructed when score standardisation was applied. Our analysis showed that we could obtain highly accurate confidence intervals for any number of sample queries while using only five standardisation systems.

## References


[1] W. G. Cochran. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 30, pages 178–191, 1934.

[2] William Feller. *An Introduction to Probability Theory and Its Applications*. Wiley publications in statistics. John Wiley and Sons, Inc., New York, 2nd edition, 1975.

[3] Sabrina Keenan, Alan F. Smeaton and Gary Keogh. The effect of pool depth on system evaluation in TREC. *Journal of the American Society for Information Science and Technology*, Volume 52, Number 7, pages 570–574, 2001.

[4] William Webber, Alistair Moffat and Justin Zobel. Score standardization for inter-collection comparison of retrieval systems. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, New York, NY, USA, 2008. ACM.