# Toward Rank Correlation as a Measure of Confidence in Information Retrieval Experiment Results

by

Alenka Bavdaz

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday August 2, 2018 at 10:00 AM.

| | | |
|---|---|---|
| Student number: | 4615093 | |
| Project duration: | January 2, 2018 – August 2, 2018 | |
| Thesis committee: | Prof. dr. A. Hanjalic, | TU Delft |
| | Dr. J. Urbano, | TU Delft, supervisor |
| | Dr. C. Hauff, | TU Delft |
| | Dr. C. C. S. Liem, | TU Delft |

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

# Abstract

In the field of Information Retrieval (IR), test collections are an important part of IR system evaluation. When evaluating IR systems on a test collection, the results may not accurately represent the performance of the systems on topics not contained in that test collection. Therefore, we want to get a sense of the accuracy of results on a given test collection. In this thesis, we use an approach that *estimates* the accuracy of test collections by estimating rank correlation between the observed and true mean scores of systems. We further evaluate this approach on new data and develop interval estimators as well. This way we provide a better sense of confidence on the system evaluation results by accounting for the inherent variability in sampling topics.

# Contents

# 1 Introduction

Information retrieval (IR) systems can be found in many aspects of our daily life, and is therefore a widely researched field. Based on predefined search terms, and the importance of the properties of the information being retrieved, an IR system returns results in order of usefulness. One of the most common forms of IR systems are search engines. Based on a given input query, we expect to receive documents containing information that is relevant to that specific query.

The relevance of the resulting documents that a system returns decides the satisfaction of a user, which is why an accurate and reliable way of evaluating the results is much needed. One of the prevalent approaches to this is to use test collections. These collections contain a corpus of documents, a set of topics, and a set of relevance judgments that indicate what documents are relevant to the topics (Sanderson, 2010). Test collections aim to form a representation that is as close to real-life usage as possible. Preferably, we would like to test the systems on all possible documents and topics that can occur. Generally, this means that test collections aim to be as large as possible, containing as many topics as possible (Voorhees and Buckley, 2002). However, creating large test collections that span a wide range of topics is expensive. And in the case of an infinite amount of possible topics, it will not cover all possible data, however large the test collection may be. Consequently, test collections should be considered to be a sample of the domain of all possible topics.

As one may expect when comparing multiple systems, the level of effectiveness of them will differ. This difference can be quantified by evaluation measures. These measures evaluate the results IR systems according to the evaluation measure criteria, on topics of the desired purpose. The results of this evaluation depend heavily not only on the topic set size, but also on which topics are used. The performance of an IR system can be represented as a distribution over topics, which is only known for the topics it has been evaluated on. Therefore, we do not know its performance on other topics.

## 1.1. Motivation

The results returned by evaluating as explained above are strongly reliant on the properties of the test collection, and don't necessarily represent the general performance of a system. This raises a few questions: *how do we know which system is really the best according to the given measure?* and *how sure are we about the reliability of an evaluation based on a given test collection?* These questions motivate research in the measurement of the reliability of a test collection.

Currently, the reliability of test collections is commonly evaluated through statistical significance testing (Smucker et al., 2007) (Cormack and Lynam, 2006). Another way to tackle the problem is to estimate the test collection accuracy, such as with the split-half method (Voorhees and Buckley, 2002).

To evaluate a system, it is run on all the topics in a test collection. The scores of which are averaged to form the *observed mean score* of the system for this test collection. If a test collection would contain all possible topics, the mean observed score would be the system's *true mean score*: the system's theoretical score when run on a test collection containing all real-life configurations.

A new method has been proposed to simulate data-sets such that the true mean score is known (Urbano, 2016; Urbano and Nagler, 2018). This way, the estimate computed by some estimator can be compared to the true mean score to gain insight on the reliability of the test collection.

In this thesis we evaluate an extension to an existing approach by Urbano and Marrero (2016) to measure

the reliability of a test collection. The goal of the extension is to get a sense of the variability the original approach may have. We hope to further reinforce the measure of reliability of any test collection evaluated with this approach.

## 1.2. Thesis outline

The thesis will be of the following structure: in Chapter 2 some background knowledge is presented, and the problem of measurement of reliability is further elaborated on. In Chapter 3, related work will be reviewed, and an overview of the foundation of this project will be given. In Chapter 4 we will go over the approach of this project, and all methods used. In Chapter 5 the evaluation method and its set-up is defined, and the results shown. The results of the evaluation will be discussed in Chapter 6, and suggestions for further research are given. Concluding statements and the contribution of this thesis can be found in Chapter 7.

# 2
# Background

## 2.1. System evaluation

To be able to see how well an IR system would satisfy real users, we want to evaluate them. There are several criteria on which evaluation measures will depend. A list of documents returned by an IR system could be unordered or ranked. The rank of a document could be taken into account when evaluating, for example: penalize a system that ranks relevant documents lowly, or the other way around. Systems may put more importance in retrieving all relevant documents, or instead most retrieved documents being relevant. Documents may also have dependencies among each other.

Based on these criteria, evaluation measures have been defined. We will explain those that are relevant to this thesis here.

### 2.1.1. Precision & Recall

The most fundamental evaluation measures are *precision* and *recall*. *Precision* is a measure that represents the number of relevant documents retrieved compared to the total number of retrieved documents:

$$Precision = \frac{\#(\text{relevant documents} \cap \text{retrieved documents})}{\#(\text{retrieved documents})}$$

*Recall* represents the number of relevant documents retrieved compared to the total number of relevant documents:

$$Recall = \frac{\#(\text{relevant documents} \cap \text{retrieved documents})}{\#(\text{relevant documents})}$$

In cases where only the top few retrieved documents matter, evaluation measures such as *P@k* can be used. This stands for *Precision at k*, which considers only the top $k$ retrieved documents, and applies *precision* as described above. In this paper, *P@10* will be used.

### 2.1.2. Average Precision (AP)

While *precision* and *recall* look at an un-ordened set of documents, *average precision* (*AP*) considers a ranked list of documents. It therefore considers documents with respect to each other. *AP* calculates *precision* for all documents above each relevant document in the ranking:

$$AP = \frac{1}{\#(\text{relevant documents})} \sum_{i=1}^{n} \text{precision of top } i \text{ documents} \cdot relevance(i)$$

where $relevance(i)$ is a binary function.

This gives us a way of evaluating high versus low ranked elements, which gives us more precise evaluation.

## 2.2. Test Collections

One approach of evaluating IR systems is to run them on test collections. These are large collections of data, containing as many as possible of the following: documents, topics, and relevance judgments. Queries are

possible search terms real users would search for, which are derived from topics (by systems). Topics are meant to abstract a user's real world need. Relevance judgments are manually recorded indicators of how relevant a document is for a topic or query. The purpose of this is to simulate running the systems on many possible queries and topics it could be presented with, to test how it would preform on real new data.

The IR systems return what they perceive as the relevant documents for a query, which are then compared to the corresponding recorded relevance judgments. This will give the system a score on how well it performed on that particular query. Here it becomes clear that repeating this for many different queries will give the best impression of how good the IR system would be at satisfying real user queries. The more queries we can run the system on, the better it reflects the system's true performance. At the same time, creating a large test collection is expensive, and running a system on a large test collection is computationally expensive. Therefore, when creating a test collection one must balance making it large enough to cover many representative topics. But small enough such that it is feasible to create one. There is an entire field of study dedicated to finding the most effective methods of selecting which documents, queries and topics to include in a test collection (Sanderson, 2010).

# 3

# Related Work

Using the evaluation measures such as the ones mentioned in the previous chapter, it is possible to quantify the performance of a system. The reliability of these results can be estimated based on several criteria. These criteria, and some measures that can be used to compare scores will be discussed in this section. First, the problem of reliability and accuracy is explained in more detail.

## 3.1. Test Collection Accuracy

Consider a researcher that wants to evaluate some systems on a test collection. We call the test collection $X$. The true mean scores of all the systems (according to some effectiveness measure) will be called $\mu$. Since we don't know the true mean scores, the goal is to estimate them as accurately as possible with test collection $X$.

If we did know the true mean scores $\mu$, we could calculate the accuracy of these estimations by comparing them to $\mu$. Thus, we define this problem as:

$$A(X, \mu)$$

where $A$ expresses accuracy as a function. The accuracy reflects how well the test collection reflects the true performance of the systems, considering the criterion of $A$.

Depending on which aspects of the test collection we are interested in, the accuracy function $A$ could consider several criteria. For example: it could consider all systems with equal importance, or put more weight on the top few. It could consider the absolute difference between system scores, or the relative difference. It could compare the value of scores, or only consider their sign (negative or positive).

## 3.2. Measures of Accuracy

We will review a few different measures that calculate the accuracy of a test collection. We can already divide the measures into two categories: *ad-hoc* (or data-based) measures, and *statistical* measures. What separates the two is the criteria on which they measure accuracy of a test collection $X$ with respect to the true mean scores $\mu$ (Urbano, 2016).

### 3.2.1. Ad-hoc measures

Ad-hoc measures look at whether systems are swapped with respect to their mean scores in the observed versus true mean scores. This gives us insight on the overall resemblance of the observed and true mean scores, and therefore the accuracy.

A way to clearly show swaps is to rank the systems according to their mean scores. Then rank correlation can be computed between the two ranked lists of observed and true mean scores. Two popular rank correlation measures that will be used in this thesis are Kendall's $\tau$ and $\tau_{AP}$. [1]

[1] Under some effectiveness measures, systems are likely to have the same mean score, so different formulations may be used. However, this goes beyond the scope of this thesis. For the treatment of these cases, the reader is referred to (Kendall, 1948; Urbano and Marrero, 2017)

#### Kendall $\tau$

In 1938, Kendall published a new measure of rank correlation (Kendall, 1938). It is a method to be able to quantify the similarity between two ranked lists (Sakai and Kando, 2008; Voorhees, 2001). In our case, the elements in these lists are IR systems. The Kendall $\tau$ rank correlation compares an element to each other element, and records +1 if the pair is in the correct/same relative order, and −1 if it is not. This is done for each element pair. The total is summed and divided by the maximum score. When pairs are in the same relative order, they are called a concordant pair, if they are swapped they are called a discordant pair. The formula for Kendall's $\tau$ is therefore:

$$\tau = \frac{\#\text{concordant} - \#\text{discordant}}{\frac{n(n-1)}{2}}$$

where $n$ is the number of the systems in the rankings. The resulting value ranges from −1 to 1, where 1 is returned when the rankings are exactly the same, and −1 when they are exactly opposite.

#### AP Correlation ($\tau_{AP}$)

Yilmaz et al. (2008) published an adaptation of Kendall's $\tau$ in 2008 (Yilmaz et al., 2008): AP Correlation - also known as $\tau_{AP}$. The purpose of this adaptation is to put more importance on higher ranked elements. Instead of comparing each element to any other elements, it only compares each element to other elements ranked above itself. The result is therefore balanced with a division of the score by $(i-1)$ for the element ranked at $i$, since there are precisely $(i-1)$ elements ranked above it. The resulting formula is as follows:

$$\tau_{AP} = \frac{2}{(n-1)} \sum_{i=2}^{n} \frac{\#\text{concordant}(i)}{(i-1)} - 1$$

where #concordant($i$) is the total number of elements above $i$ that are in the correct order with respect to $i$. $\tau_{AP}$ also ranges from −1 to 1.

### 3.2.2. Statistical Measures

One of the major reasons of the popularity of ad-hoc measures such as Kendall $\tau$ (Kendall, 1938) and AP correlation (Yilmaz et al., 2008) to evaluate systems is that they provide a clear single effectiveness score, which is easy to compare. As previously mentioned, test collections are merely an approximation of all documents, topics and true relevance judgments. Regardless of the effectiveness measure chosen, it is crucial what the resulting score means and whether it is reliable. For this reason, it is important to look further than the single effectiveness score.

Statistical measures approach this by considering the variability in the observed mean scores. In statistical hypothesis testing, a criterion is chosen on which to compare (two) systems (Box et al., 1978; Smucker et al., 2007). A null hypothesis $H_0$ is defined, which generally states that the means of the systems are the same, meaning that any difference in resulting scores is caused by random chance. Following this, a significance level $p$ that is determined based on the set criterion is used to decide whether to reject $H_0$, which occurs when $p$ is below a certain threshold. When the $H_0$ gets rejected, it provides some evidence that there is a significant difference between the evaluated systems.

As one expects, statistical tests can cause incorrect conclusions. These errors are classified in two types: Type I and Type II errors. Type I errors incorrectly reject $H_0$, and are also referred to as *false positives*. Type II errors incorrectly fail to reject $H_0$, also referred to as *false negatives*.

#### Generalizability Theory (GT)

While based on statistical theory, *genaralizability theory* (Bodoff and Li, 2007; Shavelson and Webb, 1991) does not use a null hypothesis on the difference in means of systems. Instead, it considers the variance in a behavioral measurement - in IR this is often effectiveness scores. *GT* then dissects this variance, and classifies it into different *facets*. In IR, these are characteristics such as system differences and query difficulty. Ideally, the variance would only come from the difference in systems. This way, it can indicate whether the test collection is reliable, by looking at the variance of all other characteristics.

#### Student's t-test

A popular statistical hypothesis test that is used in Information Retrieval is the *t-test*. As this measure relies on assumptions of the underlying distribution, this measure is classified as *parametric*. For this measure, the

null hypothesis $H_0$ is that the means of two systems is equal. To do this, it assumes they are random samples from the same Normal distribution. Because of this assumption, it will have more Type I errors (Sanderson, 2010). However, it has been shown to be a useful statistical significance test in IR regardless (Smucker et al., 2009).

#### Wilcoxon test

Another statistical hypothesis test is the Wilcoxon test (Wilcoxon, 1945). For this measure, the null hypthesis $H_0$ is that both systems have the same distribution. The Wilcoxon test uses the scores of each pair to calculate a difference, sorts them in an ascending order based on the (absolute) difference, and labels this difference with − for negative differences, and + for positive differences. Following this, the Wilcoxon test only utilizes this ranking, and does not look at the difference any further. This makes the Wilcoxon test more efficient computationally, but does cause of loss of information. This is why this measure is most suitable for gaining a rough impression of the difference between the two systems.

#### F-test

There are also methods to test multiple systems at the same time, such as the *F-test*. This measure analyzes the differences of means of all populations (of the systems). For this measure, the F-test statistic is used, and the results for the systems are compared. The $F$ value represents the difference between the mean values of the systems, and is also used to determine the $p$ value. It quantifies the variability of the populations of the systems using Analysis of Variance (ANOVA), further assuming that they are normally distributed.

In this case, the null hypothesis states that all systems have the same mean. This is already proven to be false if at least one system is different. Again, this does not give us a general indication of the difference *between* systems.

## 3.3. Estimation of accuracy

Since we don't know the true mean scores of the systems ($\mu$), we cannot compute the actual accuracy. Instead, we will estimate the accuracy of a test collection $\hat{A}(X,\mu)$. There are various methods to estimate this, some of which will be described in this section.

### 3.3.1. Split-half

A commonly used estimator is the *split-half* estimator, introduced by Voorhees and Buckley (2002). This method randomly samples two disjoint subsets of $n$ topics without replacement, and takes the corresponding scores from the original collection to make subsets $X'$ and $X''$. It assumes one subset corresponds to the true scores ($X''_T$), and computes the accuracy of $X'$ with respect to $X''$:

$$A(X', X''_T)$$

The expected accuracy of a random set is then the mean of the computed accuracy of many such splits $\overline{A}$. By sampling without replacement, the largest each of the samples can be is half of the number of topics $n_t$. Therefore, it repeats this process for different sample sizes. For example, with $n_t = 30$, it would split and compare two samples of 5 topics, again for 10, and the maximum 15. Based on the results, it can extrapolate the mean for $n_t$.

Sanderson and Zobel (2005) proposed to sample with replacement, such that the extrapolation would not be necessary. In this thesis, we refer to this version of the *split-half* estimator.

## 3.4. Foundation of Thesis

Statistical hypothesis testing, and statistical measures in general, make many assumptions about the data. As an example: the F-test assumption that populations have a normal distribution is not a realistic assumption in real-life data. Also, using the null hypothesis that two systems (or populations) are the same will be easily contradicted if the sample size is large enough. Topic variability can also cause random error (Cormack and Lynam, 2006). Furthermore, Rijsbergen (1979) has found that "there are no known statistical tests applicable to IR", as they are not suitable to work on IR data. We can see that correlation gives more information.

Urbano (2016) suggested a new method to estimate the test collection accuracy $\hat{A}$. The purpose of their method is to provide a general idea of how similar the observed ranking of system scores is to the true ranking of the systems, as a way to estimate the accuracy of the test collection. To do this, they use statistical estimation of the $\tau$ and $\tau_{AP}$ correlations between the observed and true mean scores. The formulas they derived for

the rank correlation measures $\tau$ and $\tau_{AP}$ are written with Bernoulli random variables called $D_{ij}$ that represent whether a pair of systems is concordant or discordant.

$$\tau = 1 - \frac{4}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_{ij}$$

$$\tau_{AP} = 1 - \frac{2}{m-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{D_{ij}}{j-1}$$

Here, it is assumed that the systems are sorted in descending order according to their observed mean score.

To calculate $E[\tau]$ and $E[\tau_{AP}]$ we need the expected value of $D_{ij}$: the probability a pair is swapped. Thus:

$$E[D_{ij}] = P(\mu_i - \mu_j < 0) = p_{ij}$$

Therefore, the correlation between the observed and true mean scores can be expressed with the number of discordant pairs. In Urbano and Marrero (2016) they further studied the probabilities of each pair being "swapped" (discordant) by estimating independently with parametric estimators *maximum likelihood* and *minimum squared quantile deviation*, and non-parametric methods *resampling* and *kernel density*. These estimators were compared to the *split-half* with replacement estimator. These estimators will be explained in Subsection 4.1.

# 4

# Approach

In this chapter we will discuss the approach taken to better measure the accuracy of test collections.

## 4.1. Point estimators

A commonly used method of estimating rank correlation is split-half estimation. However, this method is proven to be computationally expensive and produces biased estimates when used to estimate $\tau$ and $\tau_{AP}$ (Urbano, 2016). Therefore, we will instead use a statistical estimator for the two rank correlation measures as proposed by Urbano and Marrero (2016) and as explained in Section 3.4. The formulas for estimating $\tau$ and $\tau_{AP}$ are as follows:

$$E(\tau) = 1 - \frac{4}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} E[D_{ij}]$$

$$E(\tau_{AP}) = 1 - \frac{2}{m-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{E[D_{ij}]}{j-1}$$

where $D_{ij}$ is a Bernoulli random variable that equals 1 when a pair is discordant, and 0 if not. Its expected value $E[D_{ij}]$ is the probability that the pair is discordant. In other words: $P(\mu_i - \mu_j < 0)$. We call this probability $p_{ij}$.

Since we do not know the true probability that a pair is discordant, we use estimators. These will estimate this probability $p_{ij}$ for each pair.

### 4.1.1. Maximum Likelihood (ML)

The *ML* estimator assumes a Gaussian distribution for the scores of systems over topics. Then, it uses a likelihood function to maximize likelihood and estimate the $p_{ij}$s. It estimates the mean $\mu$ and standard deviation $\sigma$ based on the following formulas:

$$\hat{\mu} = \frac{1}{n} \sum X_i$$

$$s = \sqrt{\frac{1}{n-1} \sum (X_i - \overline{X})^2}$$

where $X_i$ is the difference in effectiveness between a pair of systems for topic $i$, and $\overline{X}$ is the mean of all differences in effectiveness between a pair of systems for each topic. Here, $n$ is the number of topics in the test collection.

To account for bias, the following bias correction is applied (Holtzman, 1950):

$$\hat{\sigma} = s \cdot \sqrt{\frac{(n-1)}{2}} \cdot \frac{\Gamma \frac{(n-1)}{2}}{\Gamma \frac{n}{2}}$$

This is to make sure that $E[\hat{\sigma}] = \sigma$. According to the *central limit theorem*, we know that $\overline{X}$ tends to a normal distribution. Therefore, we employ a *t-distribution* and calculate the probability of discordance with the

following formula:

$$p = P(\mu < 0) \approx T_{n-1}\Big(-\sqrt{n}\frac{\hat{\mu}}{\hat{\sigma}}\Big)$$

where $T_{n-1}$ is the *cdf* of the *t-distribution* with $n-1$ degrees of freedom.

### 4.1.2. Minimum Squared Quantile Deviation (MSQD)

Since in *ML* $\sigma$ would underestimate the dispersion of the population for small sample sizes, Urbano and Marrero (2016) suggested a new estimator. This minimizes the squared quantile deviations. In other words: it considers all quantiles of the distribution uniformly, and thus includes the tails of the distribution which may not be considered by *ML* for small sample sizes. The *MSQD* estimator uses the following estimators:

$$\hat{\mu} = \frac{1}{n}\sum X_i$$

$$\hat{\sigma} = \frac{\sqrt{2}\sum X_i \cdot \text{erf}^{-1}(2\frac{R_i}{n+1} - 1)}{2\sum \text{erf}^{-1}(2\frac{R_i}{n+1} - 1)^2}$$

where $R_i$ is the rank of topic $i$.

Since it uses the same estimates $\hat{\mu}$ and $\hat{\sigma}$ (though calculated differently), we can again employ a *t-distribution* to calculate the probability of discordance.

### 4.1.3. Resampling (RES)

Since both *ML* and *MSQD* assume a Normal distribution, the non-parametric alternative *resampling* will be used. It works as follows: we draw a sample of system scores $X_1{}^*...X_n{}^*$ from the original observations, and compute the mean for the sample $\overline{X}^*$. This is replicated $T$ times, such that the distribution of these sample means converges to the sampling distribution. The fraction of negative sample means is the probability of discordance:

$$p = P(\mu < 0) \approx \frac{1}{T}\sum I\big[\overline{X}_i^* < 0\big]$$

### 4.1.4. Kernel Density (KD)

The *kernel density* method estimates the probability of discordance by kernel smoothing. First, we try to fit a smoothed distribution to the pair of systems. If this isn't possible (for example, when it is a pair of identical pairs), we simply use the *resampling* method again. If it is possible to fit a distribution, we generate new scores based on this distribution, and calculate the mean of these new scores as sample means. Then, as with *resampling*, we count what fraction of the sample means are negative.

## 4.2. Interval estimates

When building a test collection, the rank correlation measures $\tau$ and $\tau_{AP}$ can be used to decide how many topics to contain in the test collection. Normally, the number of topics in a test collection would be decided to be when rank correlation is high enough. Urbano and Marrero (2016) showed that the point estimates are nearly unbiased, but with some degree of error. Thus, we need some sense of variability of this point estimate. We do this with interval estimators.

By adding a confidence interval, those building a test collection will be able to take into consideration not only rank correlation, but also variability when deciding on the number of topics. The confidence interval can also be used to provide more insight when evaluating systems on an existing test collection.

### 4.2.1. Calculation of confidence interval

To evaluate how good the chosen measures are at assessing the accuracy of a test collection, we will use them to make confidence intervals as extensions of the point estimates.

For a random sample $X$ and the parameter of interest $\theta$, we choose a confidence level $\gamma$. The confidence level expresses the probability of $\theta$ to be larger than the smallest value in $X$, and smaller than the largest value in $X$. In other words:

$$\gamma = P\big(\text{lower}(X) < \theta < \text{upper}(X)\big)$$

We choose the parameter of interest $\theta$ to be the correlation. Due to the *Central Limit Theorem*, we assume the sample distribution will follow a Normal distribution. Because of this assumption, we need the variance of the estimator. The standard error $se(\theta)$ for our case is defined as:

$$se(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$$

For the confidence interval we calculate:

$$\hat{\theta} \pm z \cdot \sqrt{Var(\hat{\theta})}$$

and thus for example for $\tau$:

$$\hat{\tau} \pm z \cdot \sqrt{Var(\tau)}$$

where $\tau$ is the point estimate, $z$ is the quantile of a standard normal distribution that corresponds to the confidence level. In our case, the estimation of the rank correlation ($E[\tau]$) will be our mean, and we need to find the $\pm$ range.

### 4.2.2. Variance of Kendall $\tau$ correlation

We can derive the formula for $Var(\tau)$ from the original formula for $\tau$:

$$\tau = 1 - \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D_{ij}$$

for $n$ systems.

We use the basic property $Var(aX) = a^2 Var(X)$ to get:

$$Var(\tau) = Var(1) + \left(\frac{4}{n(n-1)}\right)^2 Var\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D_{ij}\right)$$

Since we know $Var(1) = 0$, we are left with:

$$Var(\tau) = \left(\frac{4}{n(n-1)}\right)^2 Var\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D_{ij}\right)$$

Variance is computed for pairs of pairs. Therefore, we use the following general rule for $N$ random variables $Y$:

$$Var\Big(\sum_{i=1}^{N} Y_i\Big) = \sum_{i,k=1}^{N} Cov(Y_i, Y_j)$$

In our case, each $Y_i$ is a random variable $D_{ij}$. We get:

$$Var(\tau) = \left(\frac{4}{n(n-1)}\right)^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} Cov(D_{ij}, D_{kl})$$

Then expanding with the formula $Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$ to get:

$$Var(\tau) = \left(\frac{4}{n(n-1)}\right)^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \big(E[D_{ij}D_{kl}] - E[D_{ij}]E[D_{kl}]\big)$$

Since we know the expected value of $D_{ij}$ is the probability that systems $i$ and $j$ are swapped, thus: $E[D_{ij}] = P(\mu_i - \mu_j < 0) = p_{ij}$. We apply this to the formula:

$$Var(\tau) = \left(\frac{4}{n(n-1)}\right)^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \big(E[D_{ij}D_{kl}] - p_{ij}p_{kl}\big)$$

The last term we still need is $E[D_{ij}D_{kl}]$. This is the probability of both pairs of systems being swapped at the same time.

### 4.2.3. Variance of AP correlation

Again, the formula for the variance of $\tau_{AP}$ we will based on its original formula:

$$\tau_{AP} = 1 - \frac{2}{n-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{D_{ij}}{j-1}$$

Using the property $Var(aX) = a^2 Var(X)$ again:

$$Var(\tau_{AP}) = Var(1) + \left(\frac{2}{n-1}\right)^2 Var\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{D_{ij}}{j-1}\right)$$

Knowing $Var(1) = 0$, we are left with:

$$Var(\tau_{AP}) = \left(\frac{2}{n-1}\right)^2 Var\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{D_{ij}}{j-1}\right)$$

We use the general formula:

$$Var\Big(\sum_{i=1}^{N} a_i Y_i\Big) = \sum_{i,k=1}^{N} a_i a_j Cov(Y_i, Y_j)$$

to get:

$$Var(\tau_{AP}) = \left(\frac{2}{n-1}\right)^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \left(\frac{1}{(j-1)(l-1)} Cov(D_{ij}, D_{kl})\right)$$

Finally we use $Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$ to finally get:

$$Var(\tau_{AP}) = \left(\frac{2}{n-1}\right)^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \left(\frac{1}{(j-1)(l-1)} (E[D_{ij}D_{kl}] - p_{ij}p_{kl})\right)$$

### 4.2.4. Efficiency

It is clear that many combinations need to be calculated for a significant number of systems. Because of this, we will sample the number of pairs of pairs for which $E[D_{ij}D_{kl}]$ will be calculated. Since going over four loops would take too much computational power, and we will not be calculating $E[D_{ij}D_{kl}]$ for all pairs of pairs, we have to think of a smarter way of calculating $Var(\tau)$ and $Var(\tau_{AP})$. While we are sampling $E[D_{ij}D_{kl}]$, we will calculate all $E[D_{ij}]$.

All the terms within the four sums will be separated. We will call the sum of $E[D_{ij}]E[D_{kl}]$ for all pairs of pairs $X_1$, since we calculate all $E[D_{ij}]$ there will be $\frac{n(n-1)}{2}$ of these terms. Since $E[D_{ij}] = E[D_{ij}D_{ij}]$, we will calculate each $E[D_{ij}D_{kl}]$ when it is the same pair ($i = k$ and $j = l$). This term will be called $X_2$, and we will also have $\frac{n(n-1)}{2}$ of them. We will calculate a sample of $E[D_{ij}D_{kl}]$ in the case of a different pair, which will be called $X_3$. This term therefore needs to be normalized by:

$$X_3' = \frac{X_3}{\text{sample size}} \cdot \frac{n(n-1)}{2}$$

We make sure there are no identical pairs in $X_3$. (In the case of $\tau_{AP}$ each pair of pairs will be divided by its $j$ and $l$ while summed into $X_3$.) This way we can simply sum as follows:

$$X_1 + X_3' - X_2$$

and replace in the formula to get:

$$Var(\tau_{AP}) = \left(\frac{2}{n-1}\right)^2 (X_1 + X_3' - X_2)$$

## 4.3. Variance

The pairs of systems are dependent of each other. Therefore, to calculate the variance of two pairs of systems, we must also calculate $E[D_{ij}D_{kl}]$: the expectation that both pairs are discordant at the same time. We will need to adapt the estimators mentioned above to be able to calculate the covariances of dependent pairs of pairs.

### 4.3.1. Maximum Likelihood

Since we are looking at pairs of pairs of systems, we will need to employ a ***bivariate** t-distribution*. For this distribution, we need the mean and standard deviation of the difference in effectiveness of both pairs, as well as their correlation. The mean and standard deviation of the pairs are calculated the same way as in Subsection 4.1.1

### 4.3.2. MSQD

Just like for *ML*, we will employ a ***bivariate** t-distribution*. Again, we will compute the mean and standard deviation of the difference in effectiveness for both pairs of systems. We compute these with the same formulae as in Subsection 4.1.2. In addition, we also compute the correlation, as for *ML*.

### 4.3.3. Resampling

After going over all possible pairs and computing the probability of discordance for each of them, we will go over the sample of pairs of pairs. For each of them, we will do the following many times: we will sample scores and compute sample means. Then, we will record whether both pairs are discordant. Note that since the pairs are dependent, we cannot sample scores separately. Instead, we will sample topics, such that the sampled scores will be from the same topic for both pairs.

# 5

# Evaluation

Using the extended approach described in the previous chapter, we aim to provide a sense of variability of the estimators used to measure the accuracy of test collections. In this chapter, we will describe the experimental set-up, and evaluate our estimations and confidence intervals.

On each data-set, with several topic set sizes, we will estimate the $\tau$ and $\tau_{AP}$ correlations with each of our estimators. To evaluate these point estimates, we will compute their error and bias.

To get a sense of variability of the point estimates, the estimators will also estimate several intervals at different confidence levels. To evaluate the intervals, their coverage will be compared to the nominal coverage of the confidence level. The coverage is calculated by counting how many intervals contain the true mean.

## 5.1. Data

To be able to evaluate our estimation of the correlation coefficients, we need to know what the true mean scores of the systems are. This, of course, is simply not possible for any existing test collection, so we instead resort to simulation. In particular, we will follow the stochastic simulation method proposed by Urbano and Nagler (2018)[1]. In essence, this method builds a generative model $M$ of the joint distribution of system scores, such that we can simulate evaluation scores of the same systems but on new, random topics. Among others, the model $M$ contains the true distribution of each system, so we know beforehand their true mean scores and, therefore, the expected value of the simulated data. This data is realistic by fitting the model $M$ from previous existing data. The model adheres to real evaluation data by using a mixture of parametric and non-parametric techniques to fit them. In this thesis, the existing data that the models will be fit on are the TREC Web ad hoc collections from 2010 and 2011 (Clarke et al., 2010, 2011). These test collections contain 76 and 58 systems respectively.

In summary, for a given TREC run we can obtain the topic by system matrix $X$ of scores with some evaluation measure. This matrix is used to fit the model $M$, which encodes the true mean scores $\mu$. We note that this is *not* a model of the existing TREC systems, but *a* model of some hypothetical systems that behave like them. From the model we can then simulate a new matrix of scores $X'$ with a certain number of topics, and by construction we know that the expected value for some system $s$ is precisely $\mu_s$. We can estimate the correlation between the observed results $X'$ and the true scores $\mu$, that is, $\hat{A}(X')$, and because we know $\mu$ from the model we can assess how good our estimate is by comparing it to $A(X', \mu)$.

For this experiment, different numbers of topics will be used. It is crucial for the experiment to be able to see the difference in performance based on the number of topics used, as part of our evaluation is: *how many topics are ideal*?. The set sizes that will be used are 20 to 100, in increments of 20.

## 5.2. Baseline

We will compare our estimators to the widely used *split-half* estimator explained in Subsection 3.3.1. For the point estimates, this estimator samples two sets of topics with replacement from the original data, and computes the $\tau$ and $\tau_{AP}$ correlations on those topics. It does this $T$ times, such that the point estimate is the mean of the $T$ computed correlations.

[1] `https://cran.r-project.org/package=simIReff`

The interval of the point estimate is the observed interval of the $T$ computed correlations (between samples), calculated using empirical quantiles.

## 5.3. Evaluation of point estimates

To evaluate our estimates $\hat{\tau}$ and $\hat{\tau}_{AP}$, we plot their error and bias for different topic set sizes. We calculate those by comparing to the actual values $\tau$ and $\tau_{AP}$, which are known to us through the simulation. We therefore use the following formulae:

$$error(f_A) = E\big[|\hat{A}(X) - A(X,\mu)|\big]$$

$$bias(f_A) = E\big[\hat{A}(X) - A(X,\mu)\big]$$

Error is the expected absolute difference between the point estimate and the true rank correlation. Bias represents the tendency of the point estimate: if it tends to overestimate or underestimate the rank correlation. Therefore it is calculated as the expected difference between the point estimate and the true rank correlation.

Figure 5.1: Error of $\tau$ (left) and $\tau_{AP}$ (right) for all 4 data-sets

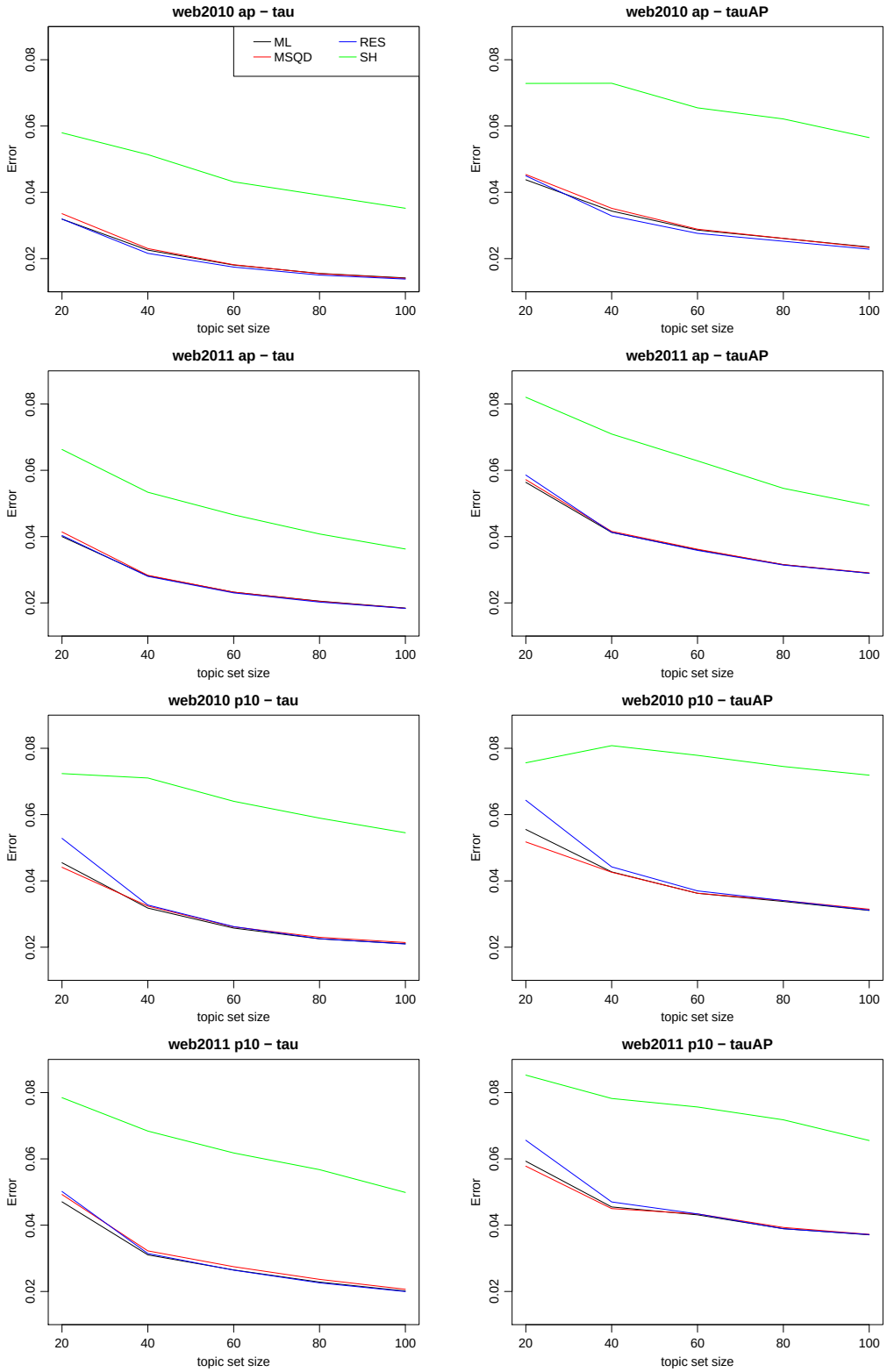
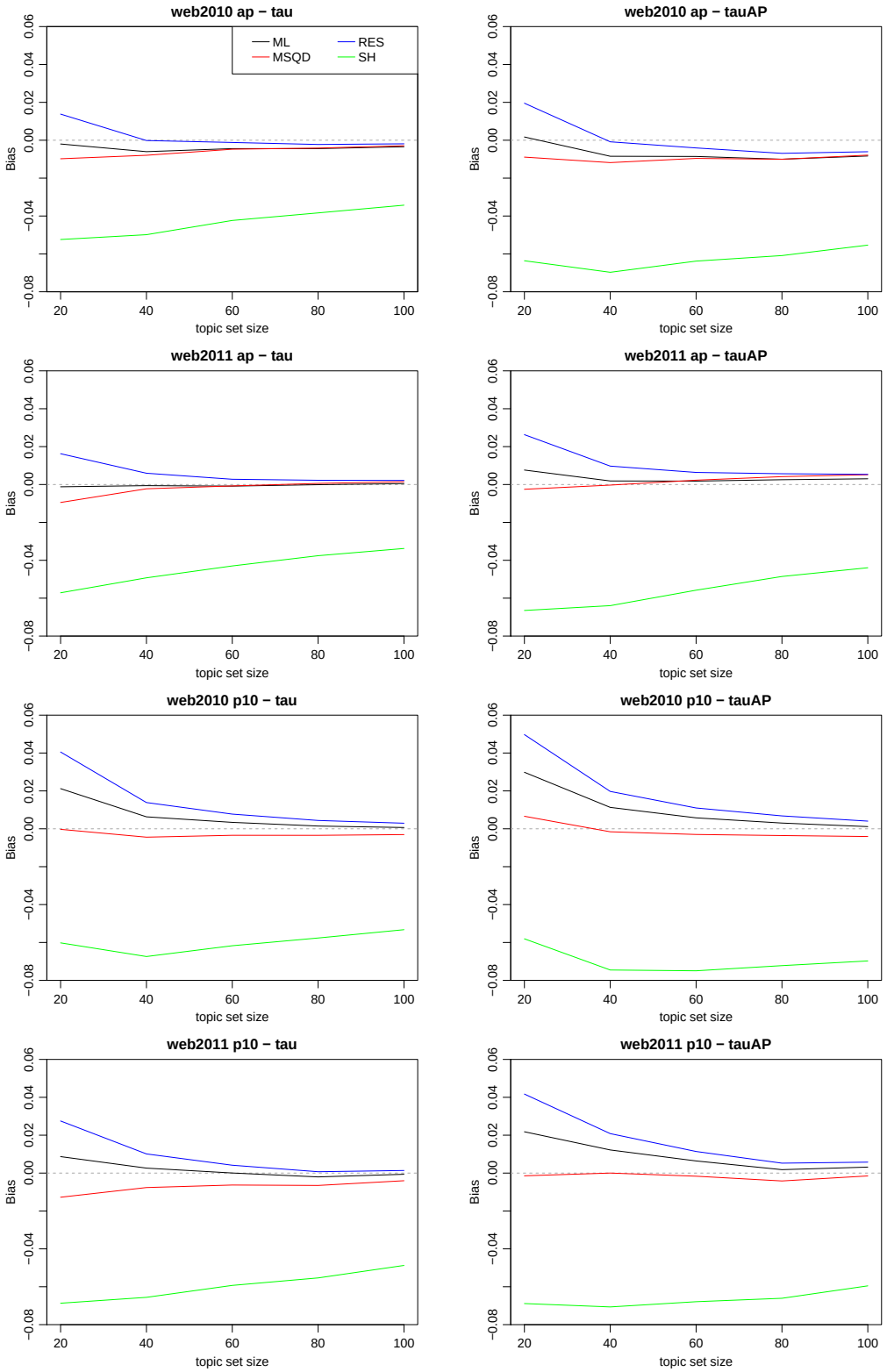


Figure 5.2: Bias of $\tau$ (left) and $\tau_{AP}$ (right) for all 4 data-sets




In Figure 5.1 we see that the *split-half* estimator always has a high error. The three proposed estimators have similar error. For the usual topic set size of 50 the three proposed estimators is around $0.02 - 0.03$ for $\tau$, and $0.03 - 0.04$ for $\tau_{AP}$. Thus, all have higher error for $\tau_{AP}$. We can see that they have higher error for smaller topic sizes. *Resampling* especially has a higher error for small topic sizes on the *P@10* data-sets. In general,

the error of all estimators is higher for the *P@10* data-sets at 50 topics.

In Figure 5.2 we see that the *split-half* estimator always has a large negative bias. The three proposed estimators are more biased for smaller topic sets. With larger topic sizes they all tend to 0 bias. We see that *resampling* always decreases from a positive bias. *Resampling* always has the highest bias, followed by *MSQD*, followed by *ML*. For the *P@10* data-sets, the bias of the estimators relative to each other is larger.

## 5.4. Evaluation of interval estimates

To compute interval estimates, we need to compute the variance. Since all pairs of pairs are dependent, we need to calculate the covariance for all of them. The total number of pairs of pairs is $\frac{n^2(n-1)^2}{4}$, which would take very long to compute. Therefore, we will instead sample a number of pairs of pairs. Since we are not calculating all, the result will be normalized accordingly as mentioned in Subsection 4.2.4.

### 5.4.1. Assumptions

Due to time constraint for this project, we will need to make some assumptions.

Within the calculation of $Var(\tau)$ and $Var(\tau_{AP})$ the only term that causes variability is $E[D_{ij}D_{kl}]$ (for different pairs, so $i \neq k$ or $j \neq l$). With respect to $\sigma_{sample}(Var)$, there should not be much difference in variability for each estimator within this term. In other words: the whole $Var(\tau)$ is larger or smaller, but the variability stays around the same. Thus, the experiment will be conducted with only one estimator.

While conducting the experiment, it became clear that the magnitude of the term described above depends most on the number of topics. While the number of systems does affect variability, its effect is not nearly as much as the number of topics.

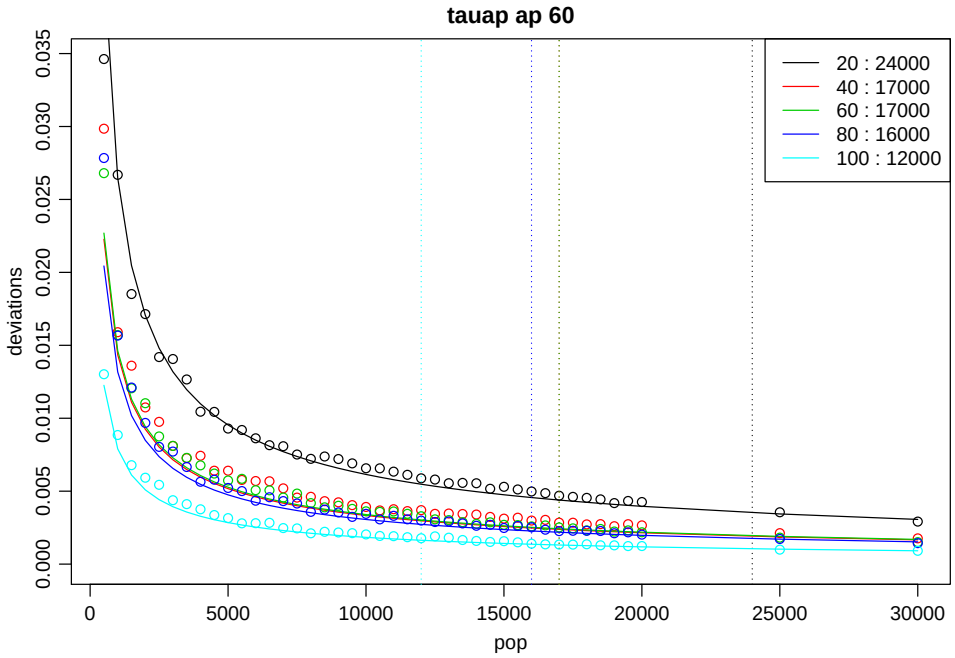
### 5.4.2. Sample size experiment

The sample size of covariance terms to be calculated in the final experiment will significantly affect the results. Therefore, it is crucial to choose an appropriate amount: large enough to get accurate results, but small enough to compute in a reasonable amount of time. We will conduct an experiment to find a satisfying number.

Using an experiment designing tool, we will compute 600 number of trials, with possible duplicates, of parameter combinations. More specifically, the effectiveness measures used in simulation, for topic set sizes 20, 40, 60, 80 and 100. All these parameters can affect the number of samples needed. A data simulation of sufficient size will be chosen to be used for all the trials, to be able to better compare variability. All estimators will run on the chosen data for 42 different sample sizes: 500 to 19500 in increments of 500, and 20000 to 30000 in increments of 5000. These numbers are chosen based on some previous experiments pointing towards the $10000-20000$ range being appropriate, therefore we want more data-points in that range. Both $Var(\tau)$ and $Var(\tau_{AP})$ will be calculated, as the rank correlation measure can also affect the sample size. This experiment will be run on all effectiveness measures.

We plot the results of all possible numbers of systems for the given combination of effectiveness measure and rank correlation method. An example of a graph can be seen in 5.3. When plotting the results, we fit a model to the data-points. The actual observations are plotted as points, and predictions based on the model are shown as lines. To decide an appropriate sample number, we draw a vertical line for each number of topics where the decrease in variability (y-axis) is smaller than $10e^{-4}$. The chosen sample size is also displayed next to each number of topics in the legend.

Figure 5.3: Sample size experiment results for AP with VartauAP and n_s=60




### 5.4.3. Interval estimates

With the variance of $\tau$ and $\tau_{AP}$, we will compute confidence intervals according to Subsection 4.2.1. To evaluate the confidence intervals around our point estimates $\hat{\tau}$ and $\hat{\tau}_{AP}$ we will compute the fraction of intervals that contain the true correlation. Then we can compare this actual coverage to the nominal coverage of the chosen confidence level, and by this evaluate how well it measures the variability. We do this for the confidence levels 60%, 80%, 90%, 95% and 99%.

Figure 5.4: Fraction of intervals containing the true mean for $\tau$ for all 4 estimators on the *AP*-based data-sets. The dashed lines represent the nominal coverage of each confidence level.

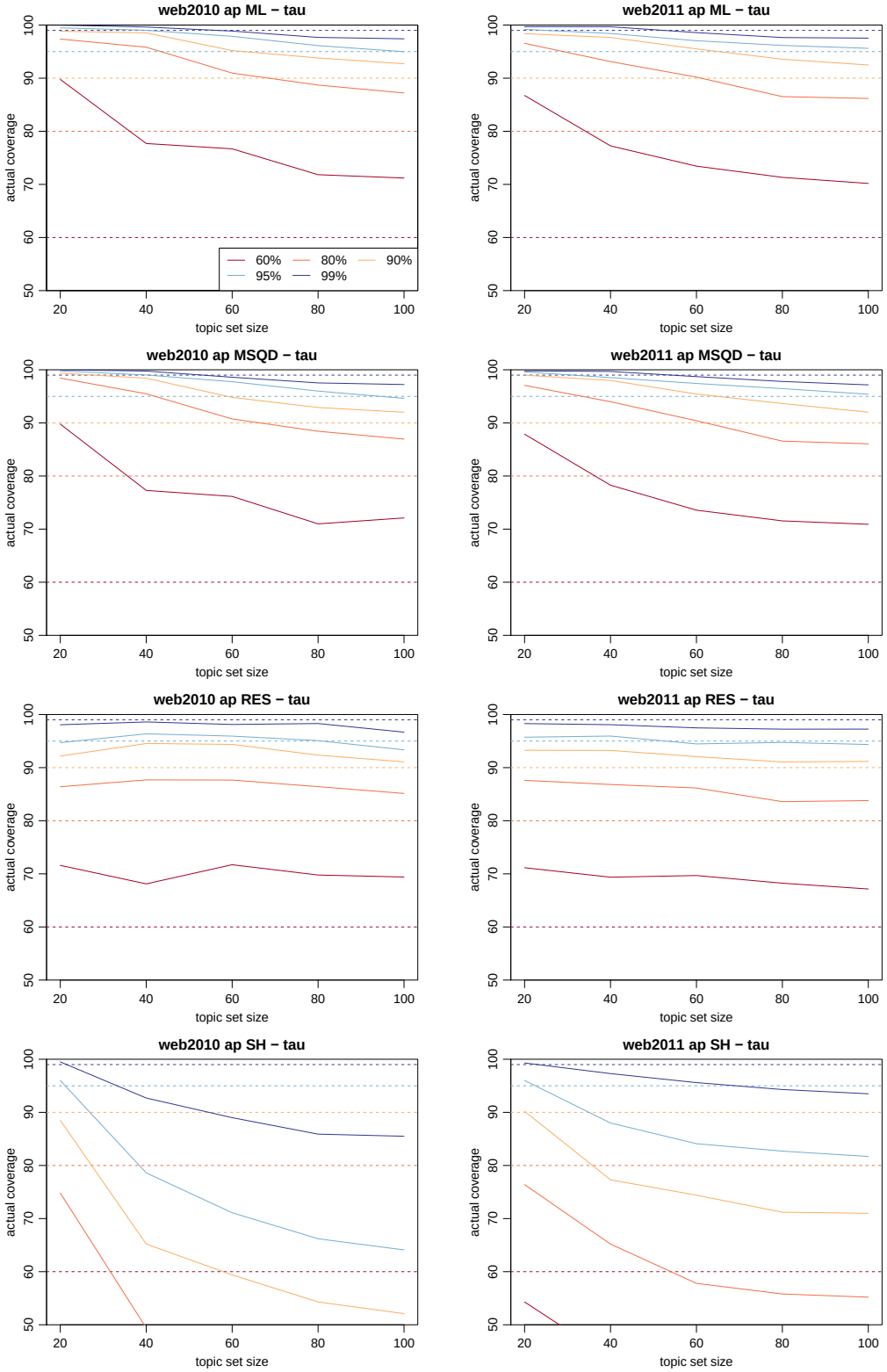


Figure 5.5: Fraction of intervals containing the true mean for $\tau$ for all 4 estimators on the $P@10$-based data-sets. The dashed lines represent the nominal coverage of each confidence level.

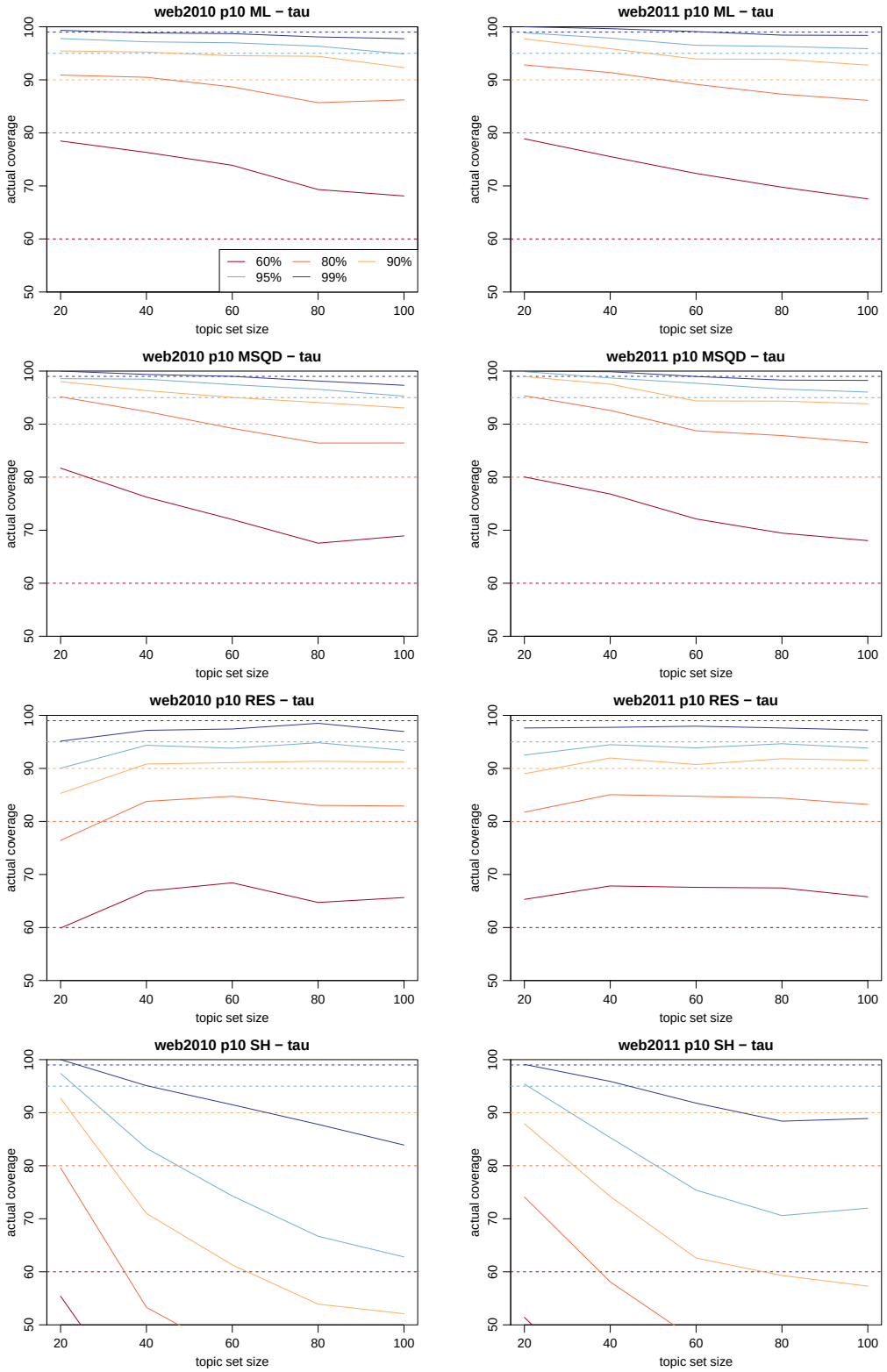


Figure 5.6: Fraction of intervals containing the true mean for $\tau_{AP}$ for all 4 estimators on the $AP$-based data-sets. The dashed lines represent the nominal coverage of each confidence level. Notice the y-axis is different for the SH estimator.

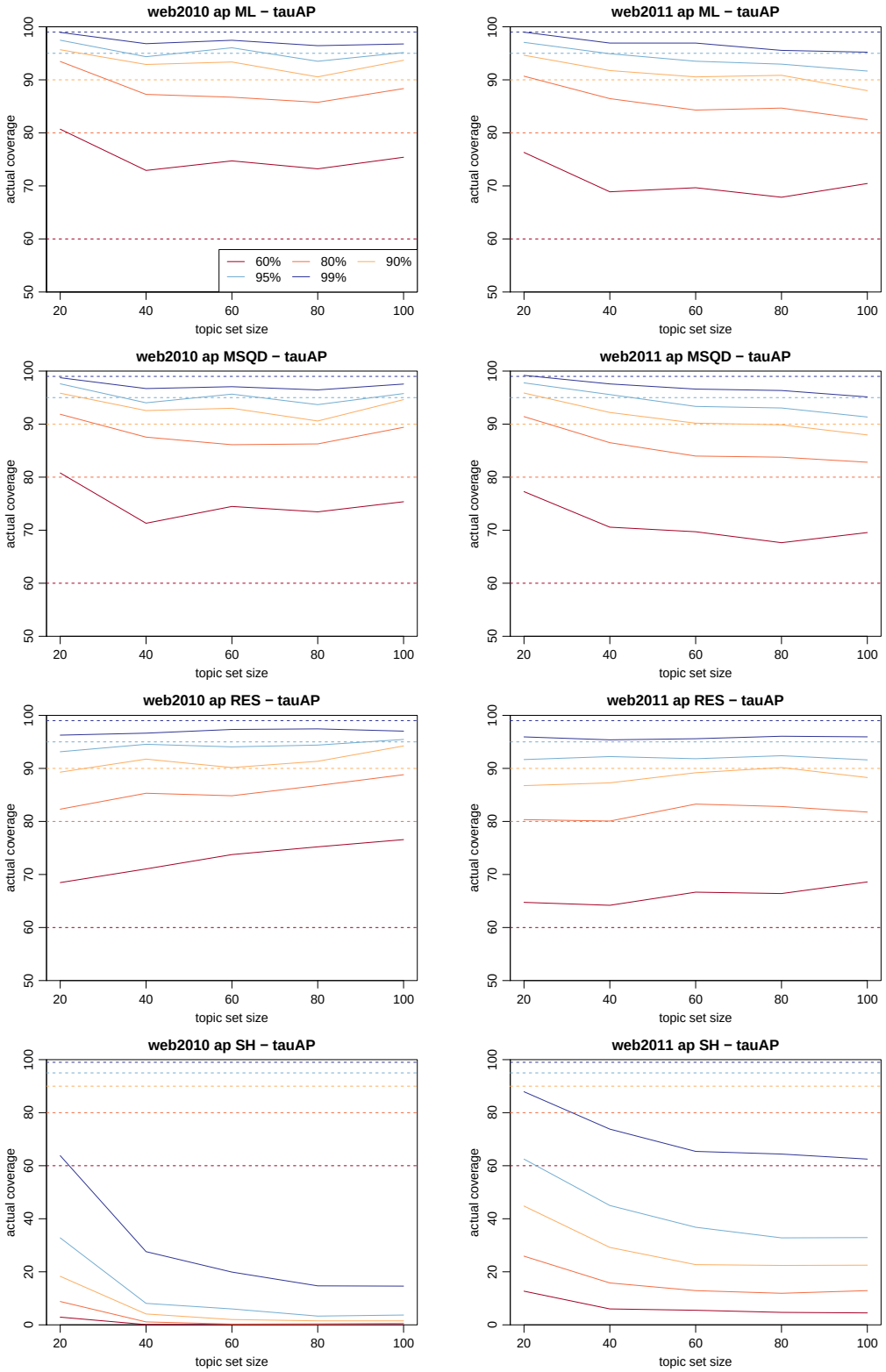
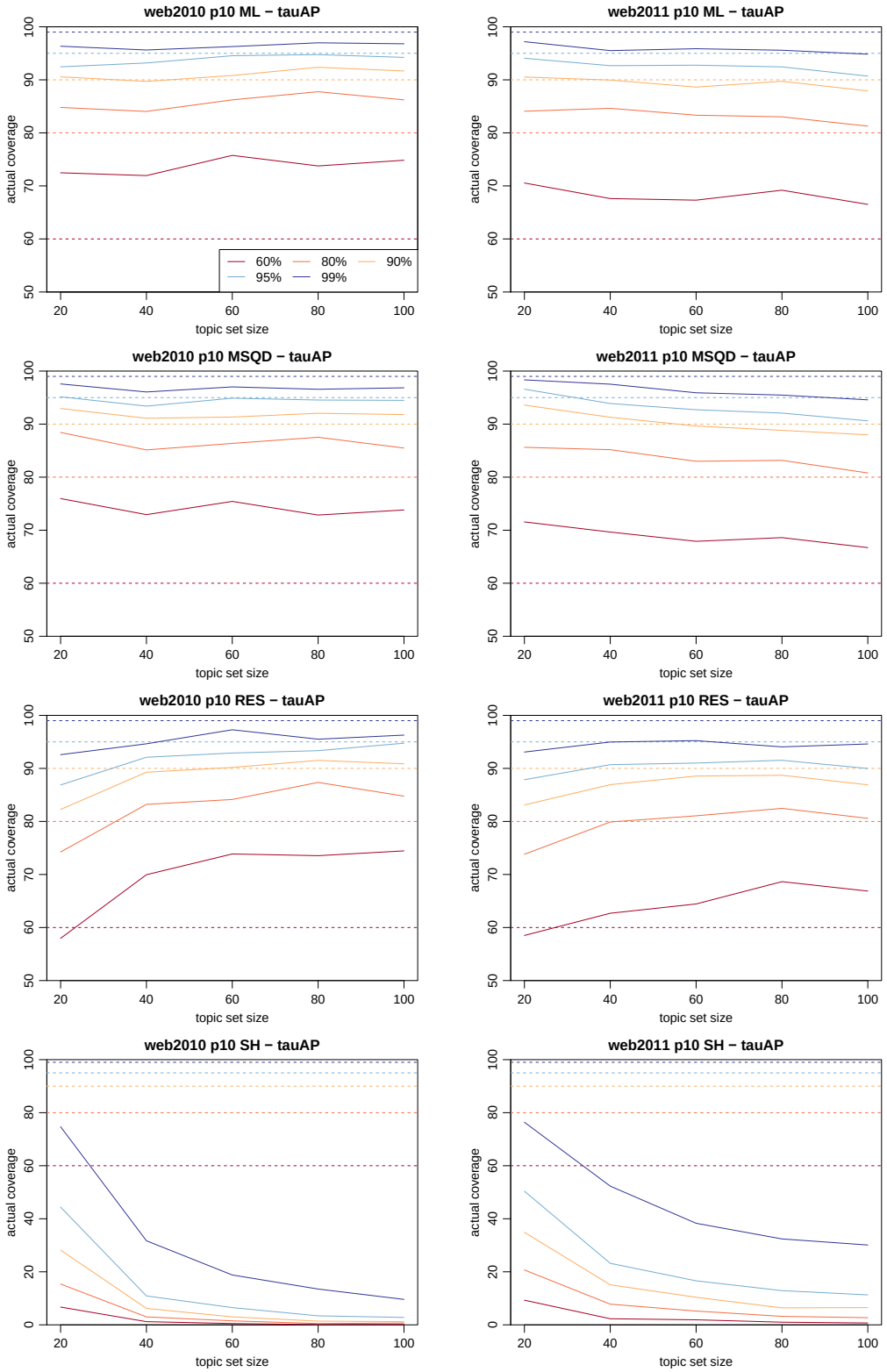


Figure 5.7: Fraction of intervals containing the true mean for $\tau_{AP}$ for all 4 estimators on the $P@10$-based data-sets. The dashed lines represent the nominal coverage of each confidence level. Notice the y-axis is different for the SH estimator.




In all figures we see that the *split-half* estimator never even comes close to the nominal coverage of the confidence levels. The proposed estimators tend to cover more than the nominal coverage for confidence levels under 99%. *Resampling* is more consistent than *ML* and *MSQD*, however seems to perform worst for

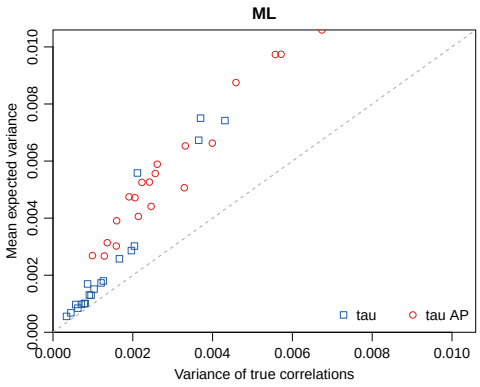
$\tau_{AP}$ on the *P@10* data-sets, as can be seen in Figure 5.7.

The coverage of *ML* and *MSQD* tend to decrease with larger topic sets in all situations, but more so for $tau$ (see Figures 5.4 and 5.5). Figure 5.6 we see that *ML* and *MSQD* seem to overestimate the most, for $\tau$ the *AP* data-sets. Otherwise, there does not seem to be a significant difference across collections or measures.

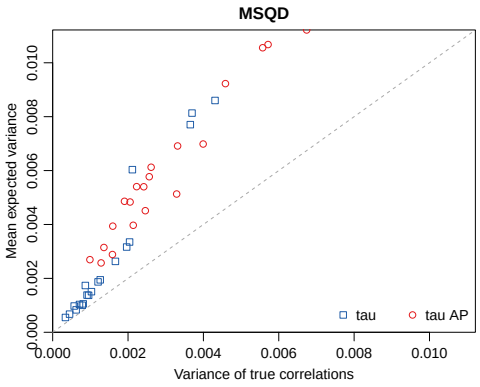
As an additional evaluation of the confidence intervals, we compare the variance of the true rank correlations $\tau$ and $\tau_{AP}$ to the mean of our computed variances for each estimator. (Considering the *split-half* estimator has already shown to be very far off from the nominal coverage, we will exclude it from this evaluation.)

Figure 5.8: Estimated vs true variability

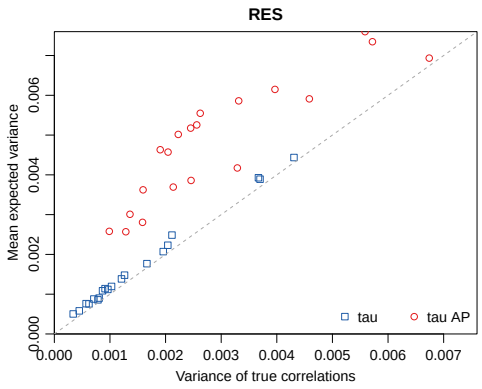
(a) Estimated vs true variability of ML




(b) Estimated vs true variability of MSQD




(c) Estimated vs true variability of Resampling




Here we see that *resampling* estimates the variance quite accurately for $\tau$. In all other cases, the estimated variance is higher than the true variance.

# 6

# Discussion

In this chapter we will discuss the results from the previous chapter. Based on the observations made, we will provide possible explanations for the behavior of the estimators. Then, some future improvements will be suggested.

## 6.1. Discussion of results

From the results, we want to derive what behavior to expect when using this approach on other test collections. As we will see in this discussion, several aspects of test collections could affect results. As an example, in this project we will see different consequences of using *Average Precision* or *P@10* in the simulation of data. However, all original test collections used are from the TREC web collections. An aspect that could effect results is the type of task of the test collection. Therefore, we will compare our results to the results of Urbano and Marrero (2016), where TREC adhoc test collections are used.

### 6.1.1. Error

We see that the proposed estimators perform similarly when it comes to error. For the usual topic size of 50, the error of the estimators being $0.02 - 0.03$ and $0.03 - 0.04$ (for $\tau$ and $\tau_{AP}$ respectively) is acceptable, as in (Urbano and Marrero, 2016). However we can see that the error still decreases to about 0.01 when increasing the topic set size to 100. This could be an argument to increase the usual topic size.

The *split-half* estimator always has a significantly higher error than all other estimators. This is likely because while *split-half* considers one half a sample and the other the truth, it does not actually compute the correlation between a sample and the truth. It actually computes the correlation between two samples.

The error is higher for the estimation of $\tau_{AP}$. This is likely due to the weight it gives higher ranked systems. Therefore, if an estimator misclassifies a swap high in the ranking, it will be penalized more, which can result in a higher error overall. We can also see that the error is higher for smaller topic set sizes. This is because there is larger variability in the rankings when there aren't many topics.

The reason *resampling* has high errors on the *P@10* datasets for small topic set sizes is likely that it does not assume a distribution. Since it does not assume a distribution, those topic set sizes are so small that *resampling* does not take into account part of the true distribution.

Compared to the results by Urbano and Marrero (2016), we obtained similar results: the *split-half* estimator has a significantly higher error, the other estimators behave similarly to each other and have approximately the same error for 50 topics, and they have higher error with small topic set sizes (such as 20). However, in their results, for small topic set sizes *resampling* generally has a higher error, below which we see *ML*, and *MSQD*, which generally has the lowest error. In our case, they are not always in that order.

Considering this, we see the result may be different for other test collections for small topic set sizes. Therefore, when building a new test collection, we recommend at least the usual 50 topics. Otherwise, these estimators will behave roughly the same as in our results.

### 6.1.2. Bias

Just as the error, the bias is much larger for the *split-half* estimator compared to the other estimators. This again is due to the reason that is explained above. Its bias is negative, which means the estimator overesti-

mates the number of swaps in the ranking.
All other estimators are more biased for smaller topic set sizes, again caused by the variability of the ranking.

We can see that the bias of the three estimators generally tends to 0. Also, the bias of each estimator still becomes significantly smaller after the usual 50 topics, which again suggests to increase it.

In our results, there is no estimator that is strictly better or worse than another. In the results of Urbano and Marrero (2016) *resampling* always had a larger bias than the other proposed estimators. This could be due to *resampling* being a less suitable estimator for the adhoc test collections, likely due to it not assuming any distribution. As explained before, the distribution of means tends to Normal, which gives an advantage to *ML* and *MSQD*. Moreover, all estimators have a positive bias, unlike in our results.

However, we can see that *resampling* does always have a positive bias. This means that it is underestimating the probability of a discordance. This could be since it does not assume a distribution, and therefore may neglect part of a distribution simply because the samples did not contain it.

Considering this, we see that the test collection may significantly influence the sign of an estimator's bias. However, there is no significant difference in how large the bias is. Therefore *resampling* will likely have the largest bias on other test collections.

### 6.1.3. Confidence intervals

As we can see in Figure 5.8, the estimators generally overestimate the variance. Because of this, the confidence intervals are larger than they should be. This leads to more intervals including the true mean, and therefore estimators' coverage being higher than the nominal coverage, as we can see in all interval results (Figures 5.4, 5.5, 5.6 and 5.7). This is likely due to calculating only a sample of covariances, and so it may see a larger variance in the estimates than when calculating all.

Another reason why the confidence intervals are less accurate is that they depend on the performance of the estimator's error and bias. Since the center of the interval is the point estimate, depending on its error and bias, the confidence interval will shift entirely.

This also explains why the *split-half* estimator performs much worse than the rest: because it had such high error and large bias, the entire confidence interval shifts, which makes it unlikely to contain the true mean.

We see that the coverage of *ML* and *MSQD* is less for larger topic set sizes. This is because the distribution that it assumes is not the true distribution, and the distribution becomes more narrow with larger topic set sizes. This leads to less confidence intervals containing the true mean. Since *resampling* does not assume a distribution, it is the most consistent estimator.

The reason why *ML* and *MSQD* overestimate more for $\tau$ on the *AP* data-sets could be since, unlike the others, they assume a (Normal) distribution, and the *AP* effectiveness measure the data is based on is (near) continuous.

Since the confidence intervals are estimates, rather than their coverage being very close to the nominal coverage, we expect to get a sense of the variability of each estimator and correlation. Therefore, even though they generally overestimate the variance, they give useful insight when considered together with the error and bias.

## 6.2. Future improvements

In this project, we have only explored a small part of an approach to measure the reliability of test collections. To get a better insight of this approach as a solution to the problem of reliability, we suggest future research on this topic to focus on the following.

### 6.2.1. Estimators

We see that estimators behave differently when compared in this approach. Depending on the test collection being evaluated, another estimator may have smaller error and bias, and more accurate confidence intervals, which would give better insight on the reliability. Therefore, more estimators should be compared.

### 6.2.2. Sampling

In the calculation of $Var(\tau_{AP})$, the covariance of each pair of pairs is weighted by the ranking of the pairs. Therefore, we lose some accuracy by sampling randomly and normalizing. As discussed, this leads to overestimating the coverage of the confidence intervals. Ideally, we would compute the covariance of all pairs of pairs. Another possible improvement could be to sample systematically, according to the system rankings.

### 6.2.3. Data sets

Since estimators behave differently on different test collections, it would be beneficial to evaluate as many as possible with this method. By covering more types of test collections, we would be able to more confidently predict the expected results of this approach on any given test collection.

Specifically, using simulated data based on different effectiveness measures. The properties of the effectiveness measures can have a significant effect depending on which estimator is used, as we have seen with the *AP* and *P@10* data-sets. Therefore, trying more may yield different results, giving more insight on what kind of results to expect when using this approach on any given test collection.

### 6.2.4. Another scenario

This type of approach can also be used for a different scenario than we have explored in this thesis. Imagine instead of evaluating an existing test collection, a researcher is building a new test collection. He wants to evaluate a number of systems on this collection. To be able to ensure some level of accuracy of the evaluation, he wants know what number of topics would be suitable. For this scenario, we want to know how accurate a hypothetical test collection is expected to be.

# 7

# Conclusion

Since system evaluation in the field of information retrieval relies heavily on test collections, it is important to be able to evaluate the reliability of test collections and the results obtained when using them. Current methods that aim to solve this problem make many assumptions, which limit them in representing the reliability accurately.

In this report we provide an extension to an approach to provide insight on a test collection's reliability by comparing the observed mean scores of systems to their true mean scores. To be able to do this, the original approach uses simulation to create datasets for which the true distribution is known. Then, it can use estimators to compute point estimates for rank correlations $\hat{\tau}$ and $\hat{\tau}_{AP}$ between the observed ranking and the true ranking. To evaluate the point estimates, their error and bias can be computed by comparing them to the true $\tau$ and $\tau_{AP}$. When evaluating the bias, we expected all estimators to have a positive bias according to Urbano and Marrero (2016). However, this was not the case, suggesting the collection has significant effect on the bias of a point estimate. Another prevalent result is that the widely used *split-half* estimator performs very badly compared to the other estimators, which can be seen both in error and in bias. For large enough topic set sizes, the performance of all three proposed estimators is similar and satisfactory for both error and bias. We note that they still improve significantly with more than the usual 50 topics, suggesting a larger number may improve IR system evaluation results on test collections.

## 7.1. Contribution

Point estimates provide limited insight on the reliability of a test collection. We have seen here and in the paper by Urbano and Marrero (2016) that the estimators have some degree of error. Therefore, to go along with these point estimates, we compute the variance of the rank correlations. With this variance, we can estimate a confidence interval to express the variability of a given estimator. In general, we saw that the estimators generally overestimate the variance, and thus the coverage of their confidence intervals is higher than the nominal coverage. Most likely, this is due to computing only a sample of covariances, which are used to calculate the variance.

Along with this, we used different test collections and effectiveness measures. With this, we were able to compare the same estimators on them, and see where the results are most influenced by them. We found that the two estimators that assume a Normal distribution (*ML* and *MSQD*) overestimate most on the *AP* data-sets.

In this thesis, we have provided an extension toward a method of measuring the reliability of any test collection.

# Bibliography


David Bodoff and Pu Li. Test theory for assessing ir test collections. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 367–374, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277805. URL http://doi.acm.org/10.1145/1277741.1277805.

G.E.P. Box, W.G. Hunter, and J.S. Hunter. *Statistics for experimenters: an introduction to design, data analysis, and model building*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1978. ISBN 9780471093152.

Charles L Clarke, Soboroff Ian Craswell, Nick, and Gordon Cormack. Overview of the trec 2010 web track. Technical report, WATERLOO UNIV (ONTARIO), 2010.

Charles L Clarke, Soboroff Ian Craswell, Nick, and Ellen M. Voorhees. Overview of the trec 2011 web track. Technical report, WATERLOO UNIV (ONTARIO), 2011.

Gordon V. Cormack and Thomas R. Lynam. Statistical precision of information retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 533–540, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148262. URL http://doi.acm.org/10.1145/1148170.1148262.

Wayne H. Holtzman. The unbiased estimate of the population variance and standard deviation. *The American Journal of Psychology*, 63(4):615–617, 1950. ISSN 00029556. URL http://www.jstor.org/stable/1418879.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. ISSN 00063444. URL http://www.jstor.org/stable/2332226.

Maurice G. Kendall. *Rank Correlation Methods*. Charles Griffin & Company Limited, 4th edition, 1948.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294.

Tetsuya Sakai and Noriko Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr.*, 11(5):447–470, October 2008. ISSN 1386-4564. doi: 10.1007/s10791-008-9059-7. URL http://dx.doi.org/10.1007/s10791-008-9059-7.

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. ISSN 1554-0669. doi: 10.1561/1500000009. URL http://dx.doi.org/10.1561/1500000009.

Mark Sanderson and Justin Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 162–169, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. doi: 10.1145/1076034.1076064. URL http://doi.acm.org/10.1145/1076034.1076064.

R.J. Shavelson and N.M. Webb. *Generalizability Theory: A Primer*. Measurement Methods for the So. SAGE Publications, 1991. ISBN 9780803937451.

Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 623–632, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321528. URL http://doi.acm.org/10.1145/1321440.1321528.

Mark D. Smucker, James Allan, and Ben Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 630–631, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572050. URL http://doi.acm.org/10.1145/1571941.1572050.

Julián Urbano. Test collection reliability: A study of bias and robustness to statistical assumptions via stochastic simulation. *Inf. Retr.*, 19(3):313–350, June 2016. ISSN 1386-4564. doi: 10.1007/s10791-015-9274-y. URL http://dx.doi.org/10.1007/s10791-015-9274-y.

Julián Urbano and Mónica Marrero. Toward estimating the rank correlation between the test collection results and the true system performance. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1033–1036, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4069-4. doi: 10.1145/2911451.2914752. URL http://doi.acm.org/10.1145/2911451.2914752.

Julián Urbano and Mónica Marrero. The Treatment of Ties in AP Correlation. In *ACM SIGIR International Conference on the Theory of Information Retrieval*, pages 321–324, 2017.

Julián Urbano and Thomas Nagler. Stochastic simulation of test collections: Evaluation scores. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 695–704, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210043. URL http://doi.acm.org/10.1145/3209978.3210043.

Ellen M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 74–82, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. doi: 10.1145/383952.383963. URL http://doi.acm.org/10.1145/383952.383963.

Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 316–323, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi: 10.1145/564376.564432. URL http://doi.acm.org/10.1145/564376.564432.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 587–594, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390435. URL http://doi.acm.org/10.1145/1390334.1390435.