Unbiased Low-Variance Estimators for Precision and Related Information Retrieval Effectiveness Measures

Gordon V. Cormack
University of Waterloo

Maura R. Grossman
University of Waterloo

ABSTRACT

This work describes an estimator from which unbiased measurements of precision, rank-biased precision, and cumulative gain may be derived from a uniform or non-uniform sample of relevance assessments. Adversarial testing supports the theory that our estimator yields unbiased low-variance measurements from sparse samples, even when used to measure results that are qualitatively different from those returned by known information retrieval methods. Our results suggest that test collections using sampling to select documents for relevance assessment yield more accurate measurements than test collections using pooling, especially for the results of retrieval methods not contributing to the pool.

ACM Reference Format:
Gordon V. Cormack and Maura R. Grossman. 2019. Unbiased Low-Variance Estimators for Precision and Related Information Retrieval Effectiveness Measures. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331355

1 INTRODUCTION

The thesis of this work is that information retrieval (IR) test collections [5] should use relevance assessments for documents selected by statistical sampling, not *depth-k pooling* or dynamic pooling methods like *hedge* [1]. To this end, we define the *dynamic* (*dyn*) estimator[1] for precision at cutoff (P@k), which is easily generalized to rank-biased precision (RBP) and discounted cumulative gain (DCG). In contrast to the well-known *inferred* (*inf*) and *extended inferred* (*xinf*) estimators [8], *dyn* is unbiased. In comparison to the *statistical* (*stat*) estimators [4], *dyn* has substantially lower variance for a given assessment budget.

When used to estimate *mean* P@k (MP@k) and other measures over a set of topics interpreted as a sample of a larger population of topics, the *dyn* estimator, coupled with an amenable sampling strategy, and a larger sample of topics, can achieve lower variance than pooling methods, for a given assessment budget.

dyn is a Horvitz-Thompson estimator [3] for the difference between the true value of P@k and a learned prior estimate. In the special case where the prior estimate is fixed at 0, *dyn* P@k is equivalent to *stat* P@k, which is also unbiased but, according to the theory underlying this work, higher variance.

[1] *See* open source `dyn_eval` implementation at `cormack.uwaterloo.ca/sample`.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331355

inf P@k and *xinf* P@k, in contrast, compute a separate estimate for each stratum of sampled assessments. However, no such estimate is possible whenever a retrieval result to be measured contains none of the assessed documents from a particular stratum. To resolve this singularity, *inf* uses Lindstone smoothing, which effectively substitutes a default constant value c when the estimate would otherwise be $\frac{0}{0}$. The published descriptions of *inf* and *xinf* use $c = \frac{1}{2}$; the reference implementation used for this study uses $c = \frac{1}{3}$. The net effect is that *inf* measurements are biased toward the value c, with small values of P@k being overestimated, and large values underestimated.

The theory that an estimator is unbiased may be falsified by identifying any possible combination of topics, documents, relevance assessment sample, and retrieval results, for which it yields a biased estimate. The experiment described here takes as ground truth the documents, topics, and relevance assessments from the TREC 8 Ad Hoc test collection [6], assuming the topics to be a random sample drawn from a population of topics with precisely the same mean and sample variance, and the assessments to be complete and infallible. To measure bias, we consider two sets of retrieval results: the 129 "TREC runs" submitted to TREC for evaluation, and 129 "dual runs" engineered to have precisely the same true P@k as the TREC runs, in 1-1 correspondence. The dual runs were formed by randomly permuting the ranks of the relevant documents in each run, while preserving the ranks of non-relevant documents.

The results of this adversarial testing show no bias for *dyn* or *stat*, small but significant bias for *xinf* (which subsumes *inf*), and very large bias, both within and between the TREC and dual runs, for depth-k pooling and hedge. To address the argument that biased measurements do not matter as long as they accurately rank the relative effectiveness of the runs, we calculate median τ correlation between the rankings achieved by repeated measurements of MP@k versus ground truth, for the TREC runs, the dual runs, and their union.

2 MEASURING P@K

Given a document $d \in D$ and a topic $t \in T$, binary relevance $\text{rel}(d) = 1$ indicates that an infallible assessor would judge d relevant to t; $\text{rel}(d) = 0$ indicates otherwise. Given a ranked list of documents $r = r_1 r_2 \cdots r_n$ from D and a topic t, our concern is how best to measure $\text{P@k} = \frac{1}{k} \sum_{i=1}^{\min(k,n)} \text{rel}(r_i)$, understanding that the infallible assessor is a hypothetical entity whose judgments can at best be approximated by real assessors, under controlled conditions, rendering $\widehat{\text{rel}}(d)$ for a subset J of all possible d. In this work, we assume the fiction that $\widehat{\text{rel}}(d) = \begin{cases} \text{rel}(d) & (d \in J) \\ 0 & (d \notin J) \end{cases}$.

If J is a statistical sample of D drawn without replacement such that each $d \in D$ is drawn with prior probability $\pi(d) = \Pr[d \in J] >$

0, we have the unbiased *stat* estimator:

$$\text{stat P@k} = \frac{1}{k} \sum_{i=1}^{\min(k,n)} \frac{\widehat{\text{rel}}(r_i)}{\pi(r_i)} .$$

The *dyn* estimator harnesses a model $\mathcal{M}(d)$ estimating $\Pr[rel(d) = 1]$:

$$\text{dyn P@k} = \frac{1}{k} \sum_{i=1}^{\min(k,n)} \mathcal{M}(r_i) + \frac{1}{\pi(r_i)} \cdot \begin{cases} \widehat{\text{rel}}(r_i) - \mathcal{M}(r_i) & (r_i \in J) \\ 0 & (r_i \notin J) \end{cases} .$$

Provided $\mathcal{M}(d)$ is independent of the outcome $d \in J$, dyn P@k is unbiased. If J is a stratified sample drawn without replacement, this constraint is met when $\mathcal{M}(d)$ is derived from $\{\text{rel}(d') | d' \in J \setminus \text{strat}(d)\}$, where strat($d$) is the stratum from which d is drawn.

A given IR method yields a different ranking $r(t)$, with a particular P@k, denoted P@k(t), for any given topic $t \in \mathcal{T}$. MP@k = $\mathbb{E}$[P@k($\mathcal{T}$)] quantifies the effectiveness of the method. Given a uniform random sample T from $\mathcal{T}$, the *dyn* unbiased estimate of MP@k is

$$\text{dyn MP@k} = \frac{1}{|T|} \sum_{t \in T} \text{dyn P@k}(t) .$$

3 STRATIFIED SAMPLING

The simplest sampling strategy that we consider divides J into equal-sized strata with equal sampling rates. For this strategy, $\mathcal{M}$ is learned using cross-validation, holding out each stratum in turn, and using the remaining strata for training. In the present study we used logistic regression to maximize (logit) likelihood $\mathcal{L}(d)$ over the training examples. We calibrated these estimates by adding a constant prior (log odds) p, and converting to probability:

$$\mathcal{M}(d) = \frac{1}{1 + \exp(-(p + \mathcal{L}(d)))} .$$

p was determined numerically to solve

$$\sum_d \mathcal{M}(d) = \sum \frac{\widehat{rel}(d)}{\pi(d)} = \mathbb{E} \sum_d \text{rel}(d) ,$$

so that the predicted number of relevant documents in the training examples would match the sample estimate.

4 PROPORTIONAL TO SIZE SAMPLING

Variance can be reduced using a model $\mathcal{P}(d)$ that predicts $\Pr[\text{rel}(d)]$ prior to determining $\pi(d)$, and, as nearly as possible, makes $\pi(d) \propto \mathcal{P}(d)$. $\mathcal{P}$ may be derived from any relevance-ranking method. In the present study we used reciprocal-rank fusion of the rankings submitted to TREC. We partitioned D into N strata of exponentially increasing size, so that higher-ranked documents were assigned to smaller strata. An equal number of documents n were drawn from each stratum S_i, and the exponential growth rate α calculated numerically to cover D when the size of the smallest stratum $|S_0| = s$:

$$\alpha = \min \alpha . \sum_{i=0}^{N-1} s \cdot \left\lfloor s \cdot (1+\alpha)^i \right\rfloor \geq |D|$$

$$|S_i| = \left\lfloor s \cdot (1+\alpha)^i \right\rfloor \ (i \in 0..N-2)$$

$$|S_{N-2}| = |D| - \sum_{i=0}^{N-2} |S_i| .$$

5 TESTING BIAS AND VARIANCE

Error err is the difference between a measurement est and ground truth tru. Bias b is the average of err over repeated measurements. MSE is the average of err^2. Variance $\sigma^2_{\text{err}} = \text{MSE} - \text{b}^2$. To test our claim that dyn MP@k is unbiased, ground truth need not be perfect, and the retrieval results to be measured need not be derived from real IR systems, as long as they are independent of J.

In evaluating an estimator, it is necessary to consider that MP@k is defined over a population $\mathcal{T}$, whereas dyn MP@k and other estimators are derived from a set of topics T, deemed to be a random sample of $\mathcal{T}$. The expectation, and therefore the bias of an estimate, is the same, whether we consider it to be an estimate of MP@k = $\mathbb{E}$ P@k($\mathcal{T}$) or of $\mathbb{E}$ P@k(T). On the other hand, even ground truth for T has non-zero variance σ^2_T when used as an estimator for MP@k :

$$\sigma^2_T = \frac{1}{|T|} \text{Var P@k}(\mathcal{T}) \approx \frac{1}{|T|-1} \text{Var P@k}(T) .$$

The inequality holds in expectation, and for the purposes of ground truth, we deem it equal:

$$\sigma^2_T = \frac{1}{|T|-1} \text{Var P@k}(T) .$$

The overall variance of an estimator is therefore

$$\sigma^2_{\text{est}} = \sigma^2_{\text{err}} + \sigma^2_T .$$

Similarly, RMSE depends on what is being estimated:

$$\text{RMS}_{\text{err}} = \sqrt{\text{b}^2 + \sigma^2_{\text{err}}} ,$$

$$\text{RMSE}_T = \sigma_T ,$$

$$\text{RMSE}_{\text{est}} = \sqrt{\text{b}^2 + \sigma^2_{\text{err}} + \sigma^2_T} .$$

6 EXPERIMENT

Using the TREC 8 data as ground truth, we compared the bias and variance of dyn MP@10 to competing statistical and non-statistical estimators with two different assessment budgets: 100 assessments per topic, and 400 assessments per topic. We also compared the effect of quadrupling the number of topics, as an alternative to quadrupling the per-topic assessment budget, given a larger assessment budget.

Ground truth was derived from the 86,830 relevance assessments (qrels) from the TREC 8 Ad Hoc task. For each of 50 topics, documents were selected for assessment using depth-100 pooling for 71 of the 129 runs. For each topic, we defined ground truth rel(d) = 1 if d was assessed and judged relevant; otherwise rel(d) = 0. An unbiased estimator should be able to estimate measurements with regard to this ground truth. For reference, we compared RMSE$_{\mathcal{T}}$ achieved by the estimators to ground-truth RMSE$_{\mathcal{T}}$ calculated using the 86,830 ground-truth assessments (1,737, on average, per topic).

To compute dyn, stat, and `trec_eval`[2] estimates, we used our own implementation, available for download as the `dyn_eval` toolkit,[1] which is input/output compatible with `trec_eval`, and has

[2] *See* trec.nist.gov/trec_eval/.

Estimator	Sample	Systematic Bias		RMS Summary over TREC Runs			RMSE	
		$\overline{b}$	$\sigma_{\overline{b}}$	b	σ_{err}	RMS_{err}	$\|T\|$= 50	$\|T\|$= 200
100 assessments per topic								
xinf	Uniform	-0.0041	0.0004	0.0707	0.0443	0.0834	0.0928	0.0768
stat	Uniform	0.0019	0.0010	0.0078	0.1154	0.1156	0.1226	0.0616
dyn	Uniform	-0.0006	0.0008	0.0061	0.0941	0.0943	0.1027	0.0515
xinf	PPS	0.0427	0.0001	0.0859	0.0103	0.0865	0.0955	0.0883
stat	PPS	0.0008	0.0003	0.0029	0.0311	0.0312	0.0513	0.0256
dyn	PPS	0.0002	0.0002	0.0031	0.0282	**0.0284**	**0.0496**	**0.0248**
depth-5	Non-random	-0.0258	0.0000	0.0349	0.0000	0.0349	0.0535	0.0403
hedge	Non-random	-0.0296	0.0000	0.0330	0.0000	0.0330	0.0523	0.0386
400 assessments per topic								
xinf	Uniform	0.0017	0.0003	0.0113	0.0365	0.0382	0.0558	0.0294
stat	Uniform	0.0002	0.0005	0.0032	0.0514	0.0515	0.0656	0.0328
dyn	Uniform	-0.0002	0.0003	0.0027	0.0383	0.0384	0.0559	0.0280
xinf	PPS	0.0277	0.0001	0.0517	0.0065	0.0521	0.0661	0.0556
stat	PPS	0.0002	0.0001	0.0011	0.0107	0.0107	0.0420	0.0209
dyn	PPS	0.0000	0.0001	0.0007	0.0082	0.0082	0.0415	0.0206
depth-20	Non-random	-0.0011	0.0000	0.0034	0.0000	0.0034	0.0408	0.0205
hedge	Non-random	-0.0014	0.0000	0.0029	0.0000	**0.0029**	**0.0407**	**0.0204**
1737 assessments per topic								
TREC	Exhaustive	0.0000	0.0000	0.0000	0.0000	**0.0000**	**0.0402**	**0.0202**

Table 1: TREC Runs

Estimator	Sample	Systematic Bias		RMS Summary over Dual Runs			RMSE	
		$\overline{b}$	$\sigma_{\overline{b}}$	b	σ_{err}	RMS_{err}	$\|T\|$= 50	$\|T\|$= 200
100 assessments per topic								
xinf	Uniform	-0.0043	0.0004	0.0709	0.0454	0.0842	0.0935	0.0772
stat	Uniform	-0.0001	0.0010	0.0112	0.1163	0.1168	0.1237	0.0626
dyn	Uniform	-0.0018	0.0009	0.0093	0.0997	0.1001	0.1081	0.0546
xinf	PPS	-0.0653	0.0001	0.1311	0.0116	0.1316	0.1377	0.1327
stat	PPS	0.0018	0.0006	0.0071	0.0734	0.0737	0.0842	0.0425
dyn	PPS	0.0008	0.0004	0.0045	0.0465	**0.0468**	**0.0619**	**0.0311**
depth-5	Non-random	-0.2202	0.0000	0.2381	0.0000	0.2381	0.2415	0.2389
hedge	Non-random	-0.1277	0.0000	0.1375	0.0000	0.1375	0.1434	0.1390
400 assessments per topic								
xinf	Uniform	0.0006	0.0003	0.0114	0.0356	0.0374	0.0552	0.0292
stat	Uniform	-0.0020	0.0004	0.0049	0.0510	0.0512	0.0654	0.0329
dyn	Uniform	-0.0012	0.0004	0.0036	0.0405	0.0407	0.0575	0.0288
xinf	PPS	-0.0261	0.0001	0.0698	0.0092	0.0704	0.0813	0.0728
stat	PPS	-0.0006	0.0002	0.0027	0.0264	0.0266	0.0485	0.0243
dyn	PPS	-0.0003	0.0001	0.0016	0.0162	**0.0163**	**0.0438**	**0.0218**
depth-20	Non-random	-0.1088	0.0000	0.1196	0.0000	0.1196	0.1263	0.1213
hedge	Non-random	-0.0526	0.0000	0.0574	0.0000	0.0574	0.0703	0.0609
1737 assessments per topic								
TREC	Exhaustive	0.0000	0.0000	0.0000	0.0000	**0.0000**	**0.0402**	**0.0202**

Table 2: Dual Runs

been verified to produce identical results. To render xinf estimates, we used the reference implementation from TREC.[3]

For statistical estimation, we considered two sampling strategies: equal-probability sampling, and probability proportional-to-size sampling (PPS) with 20 strata, as described in Section 4. For non-statistical estimation, we considered two strategies: depth-k pooling

[3] *See* trec.nist.gov/data/clinical/sample_eval.pl.

Measure	$\bar{b}$	dyn $\sigma_{\bar{b}}$	RMSE	TREC RMSE
MRBP (ϕ = 0.9)	0.0000	0.0001	0.0360	0.0351
MAP	0.0015	0.0000	0.0283	0.0280
NDCG	0.0032	0.0001	0.0323	0.0312

Table 3: dyn MRBP, MAP, and NDCG using PSS and 400 assessments per topic, compared to TREC ground truth.

and *hedge*, a best-of-breed dynamic pooling method. For depth-k pooling, we were not able to enforce a strict budget of 100 or 400 assessments per topic; as proxies we used $k = 5$ and $k = 20$, with 110 and 385 assessments per topic, on average.

For each of the 129 TREC runs and each of 129 dual runs, we measured MP@10 100 times using each statistical estimation strategy, calculating bias, variance, and error as described in Section 5 above. For the non-statistical strategies, we measured MP@10 only once, since $\sigma_{err} = 0$.

Results over sets of runs are summarized by:

- $\bar{b}$: The mean bias over all runs, an indicator of systematic bias;
- $\sigma_{\bar{b}}$: the (im)precision of the systematic bias estimate;
- RMS b: run-specific bias, over and above systematic bias;
- RMS σ_{err} : the (im)precision of the P@10(T) estimate;
- RMS err: the (in)accuracy of the P@10(T) estimate;
- RMSE_{est}: the (in) accuracy of the P@10($\mathcal{T}$) estimate.

Tables 1 and 2 shows these summary statistics for the TREC and dual runs. The results before the top break in each table use a budget of 100 assessments per topic; the results after the break use a budget of 400 assessments per topic, and the TREC gold standard uses 1,737 assessments per topic.

We see that, as predicted, the stat and dyn estimates are unbiased, while xinf shows small but significant bias, and the non-statistical methods show substantial bias in favour of the TREC runs. PPS improves the dyn and stat estimates, but harms xinf. For a total assessment budget of 20,000 documents, which is at the low end of typical TREC efforts, a budget of 100 assessments per topic for 200 topics allows dyn to surpass the accuracy of exhaustive assessment for 50 topics, for less than a quarter of the assessment effort. A budget of 400 assessments per topic for 50 topics, yields insubstantially different accuracy from exhaustive assessment. A budget of 400 assessments per topic for 200 topics yields insubstantially different accuracy from exhaustive assessment for 200 topics, with less than half the effort of exhaustive assessment for 50 topics.

7 RBP, MAP, AND NDCG

The dyn MRBP estimator is a straightforward modification to dyn MP@k. dyn MAP and dyn NDCG are somewhat biased because they divide by a normalization factor, which is also an estimate. But the normalization factor is invariant between runs, and therefore has minimal impact. Table 3 shows the results for these estimators compared to exhaustive assessment. As expected, the dyn MRBP shows no significant bias, while dyn MAP and dyn MNDCG show significant but insubstantial bias, with the net effect that all estimates have comparable accuracy to exhaustive assessment.

8 DISCUSSION AND CONCLUSION

It has been argued that bias does not matter as long as runs are properly ranked. We combined the TREC and dual runs, and ranked them using dyn MP@10, depth-20 pooling, and hedge, achieving rank correlations $\tau = 0.934$, $\tau = 0.715$, and $\tau = 0.826$, respectively. In contrast, when only the TREC runs are considered, the correlations are $\tau = 0.972$, $\tau = 0.996$, and $\tau = 0.995$. These results call into question the viability of using rank correlation over known runs as a measure of test collection accuracy. τ, like RMSE and other proposed measures, conflates bias and variance. We really have no idea whether differences in these measures reflect bias or variance.

An unbiased estimator like dyn MP@k, dyn MRBP, or dyn DCG imposes no limit on the number of topics that could be assessed, splitting the assessment budget among them. Practical considerations like the overhead of obtaining and vetting topics are likely to dominate. A minimally biased estimator like dyn MAP or dyn NDCG constrains the number of topics that may be used to a number such that its bias is insubstantial compared to variance. Even so, the optimal number of topics appears to be substantially higher than 50.

xinf, the most commonly used estimator, shows significant bias, occasioning substantial effort to discover amenable sampling strategies [7]. stat, on the other hand, shows substantial variance. Our design of dyn was directly inspired by these estimators, adopting the "default value" (albeit learned rather than constant) of xinf, and the Horvitz-Thompson estimator of stat. Our approach to stratification and our motivating application was Dynamic Sampling [2], from which dyn derives its name. Our results suggest that dyn should yield better results than stat for this purpose.

There is no limit on the number of documents or the number of relevant documents per topic to which dyn may be applied. To keep variance reasonable, it is necessary to identify a sample space that contains substantially all of the relevant documents. Twenty years ago, depth-100 pooling was found to be adequate—if imperfect—for a collection of one-half million documents and topics with 100 or fewer relevant documents. Since that time, the number of documents and the number of relevant documents have increased, while assessment budgets have decreased. Statistical sampling offers a solution.

REFERENCES

[1] Aslam, J. A., Pavlu, V., and Savell, R. A unified model for metasearch and the efficient evaluation of retrieval systems via the hedge algorithm. In *SIGIR 2003*.
[2] Cormack, G. V., and Grossman, M. R. Beyond pooling. In *SIGIR 2018*.
[3] Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 260 (1952), 663–685.
[4] Pavlu, V., and Aslam, J. A practical sampling strategy for efficient retrieval evaluation. *Northeastern University* (2007).
[5] Sanderson, M., et al. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval 4*, 4 (2010), 247–375.
[6] Voorhees, E., and Harman, D. Overview of the eighth text retrieval conference. In *TREC 8* (1999).
[7] Voorhees, E. M. The effect of sampling strategy on inferred measures. In *SIGIR 2014*.
[8] Yilmaz, E., Kanoulas, E., and Aslam, J. A. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR 2008*.