Inferential Risk Measures in Information Retrieval

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Rodger Mark Benham

Bachelor of Computer Science (Honours), RMIT University

School of Computing Technologies
College of Science, Technology, Engineering and Maths
RMIT University

January 2023

Declaration

I certify that except where due acknowledgement has been made, this research is that of the author alone; the content of this research submission is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

In addition, I certify that this submission contains no material previously submitted for award of any qualification at any other university or institution, unless approved for a joint-award with another institution, and acknowledge that no part of this work will, in the future, be used in a submission in my name, for any other qualification in any university or other tertiary institution without the prior approval of the University, and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of any published works contained within this thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my research submission to be made available on the web, via the University's digital research repository, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Rodger Mark Benham
12 January 2023

Acknowledgments

Words cannot express my gratitude for the guidance of my supervisors J. Shane Culpepper and Alistair Moffat. Their support has been steadfast throughout the entire PhD journey.

I'm also incredibly thankful for Ben Carterette's contribution, where joint collaboration with Shane and Alistair formed the basis of Chapters 3 and 4. The RMIT faculty were instrumental in keeping the thesis on track. Thank you, Mark Sanderson, Falk Scholer, Zahir Tari, Jenny Zhang, Damiano Spina, Flora Salim, Zhifeng Bao, and Jeffrey Chan for your service. Thank you, Joel Mackenzie, Luke Gallagher, Kendall Taylor, Ruey-Cheng Chen, Xiaolu Lu, Binsheng Liu, Johanne Trippas, and Oleg Zendel, for being wonderful colleagues. I would also like to extend my sincere thanks to the anonymous reviewers who provided opportunities to grow as a researcher.

I would also like to acknowledge the support I have received for my research through the provision of an RMIT Vice-Chancellor's PhD Scholarship. The COVID pandemic significantly impacted the universities' balance sheets and altered the certainty around funding. With that, I would be remiss in not mentioning the support of accommodating employers in the latter stages of the thesis.

I want to express my profound appreciation to my family and friends for their support. My parents, Craig and Dolly, and family from the O'Connell side were lifelines throughout the endeavour.

Finally, thank you to my inspiring wife-to-be, Erin, for your love and belief in me. I am so lucky to have had your support through this journey, and I'm super excited for the next chapter in our lives!

Contents

List of Figures

List of Tables

List of Symbols

s	An information retrieval system.
S	A set of information retrieval systems.
T	A set of topics for evaluation on a corpus, overloaded to represent a statistical function to Bootstrap.
R	The count of relevant documents for a topic against a corpus, or a set of document ranks for a system-topic-rank document gain value.
N	The count of non-relevant documents for a topic against a corpus, or, a Gaussian distributed variable with mean μ and standard deviation σ.
ϕ	The rank-biased precision [135] and overlap [210] persistence parameter.
d	The expected viewing depth of a ranking in connection with ϕ.
Δ	The paired set of system-topic effectiveness score differences over T between systems s_1 and s_2.
t	A Student-t distributed variable for use with the t-test, otherwise overloaded to signify an evaluated statistic T during a Bootstrap.
t^*	Bootstrap replicates of the statistic T.
z	A variable distributed by the standard Normal distribution.
α	The probability of accepting the null hypothesis, the system effect in a statistical model, or the original risk-level parameter for risk overlays.
r	A vector of relevance judgments for a ranked list in the background section, then used as an altered risk-level parameter of α from Chapter 3 onwards.
Δ_r	The paired set of system-topic effectiveness score differences over T between systems s_1 and s_2, where $\Delta < 0$ values are multiplied by r.
β	The Beta distribution, the power of a statistical test, or the topic effect in a statistical model.
m	The number of multiple comparisons to correct for, or the number of artifact systems included in a statistical model.

$\hat{\theta}$	A boundary on a confidence interval.
i	The rank of a document in a list, or the ith system in a statistical model for α.
j	The rank of a document in a list, or the jth topic in a statistical model for β.
k	The maximum viewing depth of a ranking, or the rank effect of a document gain value in a statistical model for R.
Φ	The cumulative distribution function of a discussed statistical distribution.
B	The number of bootstrap values to generate.
Y	Generic set representing a sample of observations.
$\overline{Y}$	The arithmetic mean of the set Y.
W	The Shapiro-Wilk test statistic, or the count of MCMC warmup iterations to discard.
D	The Kolmogorov-Smirnov test statistic.
C	The count of MCMC chains.
I	The count of MCMC iterations to run.
θ	A Bayesian posterior distribution of $C(I - W)$ draws.
$\hat{R}$	An MCMC diagnostic used to examine θ chain convergence.
n_{ESS}	The effective sample size of a simulated MCMC θ.
b	An intercept term in a linear model.
λ	The Skew-Normal distribution shape parameter.
γ	A Gamma distributed variable.
π	A constant component of a variable, or the standard mathematical constant.
ξ	The location parameter of a Skew-Normal distribution.
ω	The scale parameter of a Skew-Normal distribution.

Abstract

When the effectiveness of a search ranker is evaluated, an effectiveness metric is computed by averaging how well the ranker retrieved relevant documents over a set of topics on a test dataset. If a search practitioner wishes to test a hypothesis on whether one ranking algorithm is more effective than another, traditionally a null hypothesis statistical test is used. Typically rankers are tested in pairs at a time, but often many effective alternative rankers could be used to provide background context on the likelihood of a ranker outperforming another. However, current approaches for inferential testing on the outcomes of many rankers tend to reduce statistical power, so much so that prior work has questioned whether accounting for this family-wise error yields ineffectual inferential analyses with the available IR test collections. Regardless, the demand for multiple comparison correction continues to grow in scientific venues to avoid drawing conclusions under false pretenses.

Another recent feature in IR evaluation is the improved awareness that retrieval effectiveness over topics varies substantially for different rankers, where a challenger system that performs better on average may be risky to replace a champion system with due to outliers. As seminal works in economic theory show that people are more sensitive to losses than gains, they may perceive that another system selected for improved mean effectiveness is less effective overall if it previously returned effective rankings and now does not. Risk overlays aim to support replacing rankers with the joint goal of net improvement without unacceptable drops in effectiveness. Current inferential risk overlays evaluate pairs of systems, which this thesis extends towards a multiple-system testing approach.

The thesis begins by investigating the shape of risk-adjusted score distributions, with the insight that their skewness may violate the parametric assumptions of common inferential risk testing approaches. Secondly, how amenable Bayesian inference is to the problem of multiple comparison correction is explored for the first time in an IR context, factoring in the above need to handle skewness appropriately in the case of risk analysis. A novel Bayesian hierarchical modeling approach is applied to IR scores, combining the ability to use many systems as background information and the skewness properties of risk-adjusted scores. Finally, in finding that directly modeling risk-adjusted scores resulted in low statistical power, the thesis explores modeling IR scores directly and evaluating risk as an inferential summary statistic on the posterior predictive distribution. The results indicate that this method improves the discriminative capacity of risk inference over many systems while retaining the corrective properties of Bayesian hierarchical modeling.

1

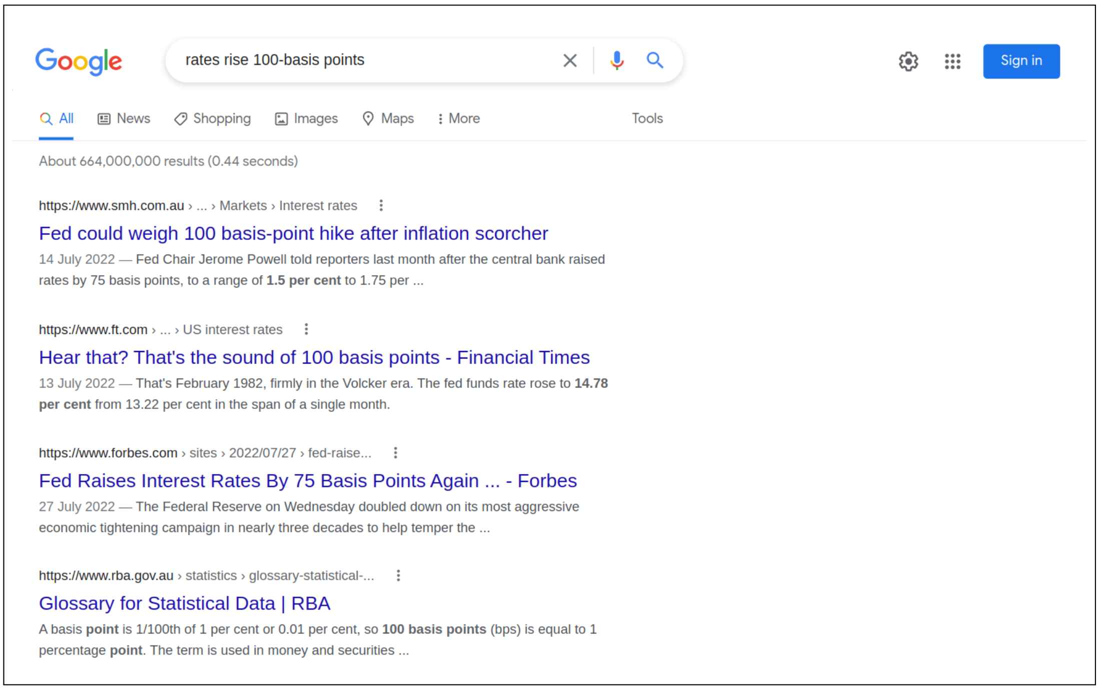
Introduction

Suppose you are in Australia and just heard on the radio that interest rates were likely to rise 100-basis points. You wonder why interest rates are being raised by this amount. So you go to your computer and launch a web browser pointing to a web search service to learn more. Google is your default search provider, so you input the *query* `rates rise 100-basis points` into the search box and submit the request to the web search service.

Figure 1.1 on the next page shows the search engine result page (*SERP*) for your *information need* after the page loads. A set of webpages are returned ordered by their predicted *relevance* to the query. For example, the first search result presented is an article published by an Australian news source, the Sydney Morning Herald, and the second is an American news article published by the Financial Times. Both articles are topically related to your query. However, the Sydney Morning Herald article may be more relevant to an Australian searcher, so it appears at the head of the list. You read the article, finding that it only discusses a US Federal Reserve interest rate rise and provides no commentary on the effect it might have on the Australian economy.

Although the webpages retrieved in the SERP shown in Figure 1.1 on the next page match the keywords in your submitted query, you question whether Google should remain your default web search provider. You wonder whether an alternative provider retrieves more useful documents for an Australian audience, as it is your understanding that Australian interest rate decisions are made by the Reserve Bank of Australia (RBA). You decide to compare two search providers on the query, where Google is the *champion* and the alternative service is the *challenger*. Figure 1.2 on page 5 shows a side-by-side comparison of the Google page rankings (shown in Figure 1.1) for the query against a SERP from a challenger.[1] On the surface, the Google SERP appears to be a more useful ranking as it ranks a news article from an Australian source at the head of the list, compared to the challenger ranker, which provides an educational article defining what basis points are. However, the most crucial aspect is whether the contents of the retrieved pages satisfy the information need. To decide which search engine produces the most effective ranking, you open the top-4 documents of each

[1] The alternative service does not actually exist yet, think of it as being what a newly-graduated PhD student might build if they won TattsLotto.



rates rise 100-basis points

About 664,000,000 results (0.44 seconds)

https://www.smh.com.au › ... › Markets › Interest rates

Fed could weigh 100 basis-point hike after inflation scorcher

14 July 2022 — Fed Chair Jerome Powell told reporters last month after the central bank raised rates by 75 basis points, to a range of **1.5 per cent** to 1.75 per ...

https://www.ft.com › ... › US interest rates

Hear that? That's the sound of 100 basis points - Financial Times

13 July 2022 — That's February 1982, firmly in the Volcker era. The fed funds rate rose to **14.78 per cent** from 13.22 per cent in the span of a single month.

https://www.forbes.com › sites › 2022/07/27 › fed-raise...

Fed Raises Interest Rates By 75 Basis Points Again ... - Forbes

27 July 2022 — The Federal Reserve on Wednesday doubled down on its most aggressive economic tightening campaign in nearly three decades to help temper the ...

https://www.rba.gov.au › statistics › glossary-statistical-...

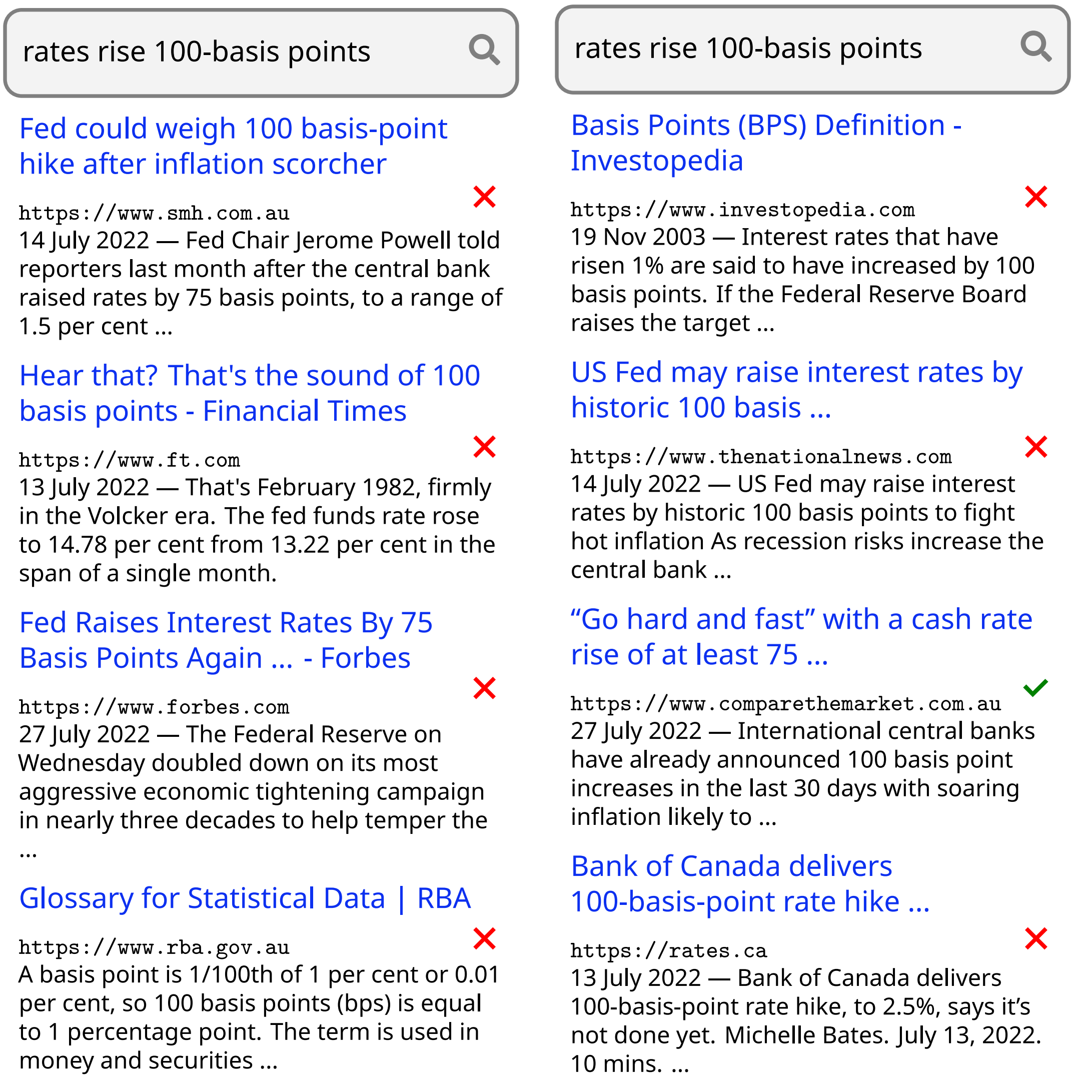
Glossary for Statistical Data | RBA

A basis **point** is 1/100th of 1 per cent or 0.01 per cent, so **100 basis points** (bps) is equal to 1 percentage **point**. The term is used in money and securities ...

Figure 1.1: *Search engine result page* (SERP) from the Google web search provider. Screenshot captured 22nd August 2022.

ranking in new tabs in the web browser, shuffle their presentation order around, and enlist the help of an impartial assessor to read each page and determine its usefulness towards the topic. 'Only pages mentioning a 100 basis point rate rise and how that might affect Australians are interesting to me', you tell the assessor. 'Too easy cob', the assessor replies, and they proceed to determine the relevance of the pages. After a stream of expletives while waiting for the National Broadband Network to serve the pages, the assessor eventually yells 'You beauty, all done!' You ask them to reveal their results. After tabulating pairs of the page titles and their associated judgments in a spreadsheet, you observe that the assessor only found the third document in the alternative ranking to be relevant.

You decide to research existing IR *effectiveness metrics* to quantitatively assess the quality of each SERP retrieved by their rankers. If the first four documents of each ranking in Figure 1.2 on the following page will always be read, then the presentation order of the results is not essential; all that matters is that the challenger ranker retrieved one relevant result out of four, making its *precision* [50] score $1/4 = 0.25$, and where Google's score is $0/4 = 0.00$. More considerations are at play when a document's position in a ranking and its degree of relevance determines retrieval effectiveness. Should the ranking be measured against the best possible ranking from the judgments identified? How severely should the score contribution of relevant documents found deeper in the ranking be penalized compared to the head of the list? There are many metrics which include rank information; you select the *reciprocal rank* (RR) metric for its simplicity. It is defined as one divided by the rank position of the first relevant document retrieved (or 0 if no relevant documents were returned in the SERP). The RR score for the Google SERP is 0, and the challenger SERP score is $1/3 = 0.33$.



rates rise 100-basis points

Fed could weigh 100 basis-point hike after inflation scorcher ✗

https://www.smh.com.au
14 July 2022 — Fed Chair Jerome Powell told reporters last month after the central bank raised rates by 75 basis points, to a range of 1.5 per cent ...

Hear that? That's the sound of 100 basis points - Financial Times ✗

https://www.ft.com
13 July 2022 — That's February 1982, firmly in the Volcker era. The fed funds rate rose to 14.78 per cent from 13.22 per cent in the span of a single month.

Fed Raises Interest Rates By 75 Basis Points Again ... - Forbes ✗

https://www.forbes.com
27 July 2022 — The Federal Reserve on Wednesday doubled down on its most aggressive economic tightening campaign in nearly three decades to help temper the ...

Glossary for Statistical Data | RBA ✗

https://www.rba.gov.au
A basis point is 1/100th of 1 per cent or 0.01 per cent, so 100 basis points (bps) is equal to 1 percentage point. The term is used in money and securities ...

rates rise 100-basis points

Basis Points (BPS) Definition - Investopedia ✗

https://www.investopedia.com
19 Nov 2003 — Interest rates that have risen 1% are said to have increased by 100 basis points. If the Federal Reserve Board raises the target ...

US Fed may raise interest rates by historic 100 basis ... ✗

https://www.thenationalnews.com
14 July 2022 — US Fed may raise interest rates by historic 100 basis points to fight hot inflation As recession risks increase the central bank ...

“Go hard and fast” with a cash rate rise of at least 75 ... ✓

https://www.comparethemarket.com.au
27 July 2022 — International central banks have already announced 100 basis point increases in the last 30 days with soaring inflation likely to ...

Bank of Canada delivers 100-basis-point rate hike ... ✗

https://rates.ca
13 July 2022 — Bank of Canada delivers 100-basis-point rate hike, to 2.5%, says it’s not done yet. Michelle Bates. July 13, 2022. 10 mins. ...

Figure 1.2: Google SERP (Figure 1.1) on the left, shown side-by-side with a ranking from an alternative provider’s SERP (right column). Sites that are relevant towards the query in an Australian context are marked with a green tick, and marked with a red cross otherwise.

You thank the assessor for their help and declare that you will change your default search provider because the alternative search engine is more effective than Google. ‘Ease up, turbo!’ the assessor says. They are not convinced that the results of one query should generalize to every possible query, and they had trouble deciding ‘relevance’ as US rate decisions strongly influence the RBA. The assessor offers to read hundreds of web pages to answer your search provider dilemma more conclusively. Your research shows that 50 topics are typically used for IR effectiveness measurement, with the average score taken. To split the assessment workload, you both define 25 new topics (that all have an Australian focus, one of them is “Crocodile Dundee”), giving $4 \times 2 \times 25 = 200$ documents to judge each. You both set aside seven hours on the weekend, assuming it takes two minutes to assess each page.

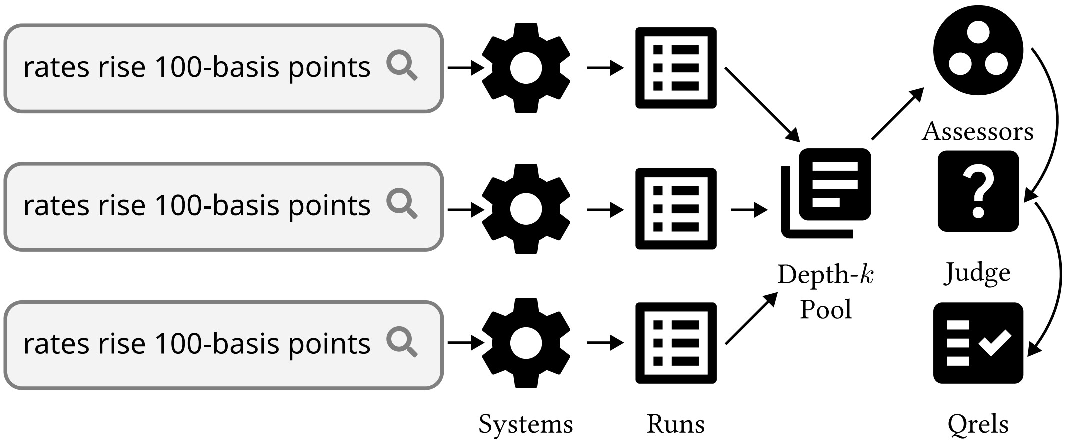

rates rise 100-basis points
rates rise 100-basis points
rates rise 100-basis points
Systems
Runs
Depth-k Pool
Assessors
Judge
Qrels

Figure 1.3: How judgments are formed for one topic in an evaluation campaign.

After collating the judgments, you calculate the mean (over the set of topics) effectiveness score of every SERP for each of the two systems. The average Google SERP score is greater than the challenger. However, general arguments about whether Google will yield greater effectiveness than the challenger on average are still unanswered, because the selected topics might have influenced the results. So you dig even deeper, finding that IR researchers use *statistical inference* to overcome the topic sample issue. A simple paired test determines whether the difference in the mean scores observed is independent of the sample of topics considered in the evaluation. You compute the test and conclude with 95% confidence that Google is the most effective ranker of the pair for the regional needs of an Australian.

In sharing the results of your investigation with other friends, you learn of their affinity towards a new Australian-based web search engine BigFactKoala that claims to be improving search effectiveness for regionally-specific needs. Rather than only considering whether one challenger search provider outperforms a champion search provider, given many challengers and one champion, which one should you select as your default search engine? You consider rerunning the same procedure involving three search systems this time, but when submitting the same queries on the original paired comparison, new documents were retrieved in practically every SERP. Figure 1.3 shows the process for acquiring judgments for one of the fifty topics. With the addition of another challenger, the number of documents to be judged could be $4 \times 3 \times 25 = 300$ documents per assessor in the worst-case scenario where every retrieved page is new, so you broach the matter with your wider friendship group.

A friend of a friend is a professor who has worked in web search evaluation for years now, and you are introduced to them. They are interested in formalizing your investigation into a research study. The professor suggests that the "Cranfield" evaluation paradigm developed in the 1960s involving a fixed document collection could be used [50], which would avoid a situation where judgments would need to be collected every time a new search engine is evaluated. As judging every document against a topic is intractable for a large collection like a web crawl, deciding a rank cut-off as was done in Figure 1.2 on page 5 is a workable approach and is formally known as pooling [109]; a concept used by the long-running Text REtrieval Conference [204]. According to the professor, the RR metric is a good choice for measuring the effectiveness of navigational topics (such as, find the RMIT University homepage), but it loses interesting details for informational topics like the one you initially investigated in Figure 1.2, as many pages could satisfy the information need. You agree with the professor and ask whether they would initiate the study, as they have the expertise and a recent offline *crawl* of the web with excellent coverage of Australian websites. They oblige, penning an open letter to the commercial search engines inviting them to produce SERPs for each of your Australian topics on the crawled offline collection. All major web search providers index the corpus and bundle each SERP for each topic into a *run* to be submitted to the evaluation campaign, interested to know where their product stands against the competition. Many open-source search systems are included to bolster the fidelity of the judgments.

Now that many businesses and academics are invested in the outcomes of this evaluation campaign, a university research center is selected to act as an impartial entity and perform the relevance assessments. Each run is pooled to depth 100, with the resultant huge set of documents checked by careful assessors for relevance to the corresponding topic [224]. With all these judgments, a team of academics computed the effectiveness scores for each run and used statistical inference to mitigate against sampling errors. They found that BigFactKoala has successfully challenged Google when it comes to satisfying the information needs of Australians. The results persuade many Australians to use BigFactKoala as their search engine when they hear about this research study on the radio and via Twitter. Seeing a spike in advertising revenue, BigFactKoala raises capital to expand into New Zealand.

Problem. After some time, BigFactKoala executives noticed that its traffic was declining. It plans to shut down and sell its services to a competitor. Despite being the most effective search engine on average over the topics considered in the Australasian region, searchers have submitted feedback explaining why they have returned to their original search provider. Many searchers had built prior workflows around the rankings provided by their original search provider for queries with many relevant answers to an information need, and the sites they expected to see were not showing up in the BigFactKoala SERP. Some would issue a query to their original search provider and follow the link rather than bookmarking pages in their browser. Others tried BigFactKoala but found it didn't work as well as the original search provider on their topics of interest.

Concerned with the hemorrhaging revenue, the BigFactKoala board of directors convene and agree to explore litigation against the professor who set out to perform the study. After all, if it were not for the outcomes of the study, the company would not have taken on so much financial risk. The directors enter their luxury vehicles and drive to the university research center to accost the professor. They encircle the professor in their windowless office; demanding an explanation for the decline of BigFactKoala, threatening to sue the university for damages.

The professor explains that the study was conducted in good faith and that the results were accurate if all search engines were being treated as equal. Although the BigFactKoala SERPs were the most effective on average, they were not always the most effective for the topics that the searchers were interested in. Voorhees and Harman [204, p. 15] notes that "users remember abject failures" and that "averages increase the reliability of the evaluation, but hide large variability in per topic effectiveness." The professor mentions that the situation BigFactKoala finds itself in was entirely predictable, and had the leaders of the company sought expert advice on web search evaluation, they would not rushed into financial ruin. After a riveting lecture on *risk-sensitive evaluation* in IR, where search practitioners weight reductions in challenger effectiveness greater than gains compared to a champion system, the directors begrudgingly accept the explanation and leave the university, forever.

A rank-and-file engineer is urgently summoned to the BigFactKoala offices to investigate how risk-sensitive evaluation could be implemented to tune the search engine and save the business. The board wants the same 95% confidence level as before, but this time they want to answer the right question. Without the ability to quantify their confidence in a new system within a multi-system risk-sensitive evaluation scenario, the board would not be able to assure their stakeholders that they can address their attrition issue in time. The engineer finds a useful paper by Dinçer et al. [67] which combines statistical inference with paired system risk comparisons, but how to perform an inferential risk analysis with multiple comparison correction when many challengers exist remains a mystery. Running out of time, money, and answers, the board makes the honorable decision to sell the business to a competitor, so that at least the staff can remain in work. Having had their fill of the rat race, the professor submits their resignation. They are rumoured to have fallen just shy of their Bahamas retirement plans, and settled on living off the Great Ocean Road, in Lorne, Victoria, Australia. After everything that happened, you are still not sure about which search engine is the most effective for your needs.

Research Goal. This thesis explores risk-sensitive approaches for evaluating the effectiveness of many search engines simultaneously with statistical assurances. Whether the risk transformation has implications for the kind of statistical tests typically used is first explored from the point of view of pairs of systems. After understanding how the risk transformation affects inferential decisions, novel methods and enhancements are made to compare many systems simultaneously with and without risk inferentially.

1.1 Contributions

1.1.1 Extending Paired Inferential Risk Overlays

Chapter 3 starting on page 59 examines for the first time the distributional properties of risk-adjusted score differences in an IR context where the goal is to use these values for statistical inference. Analyses from experiments run on two corpora and evaluation metrics indicate that deviations from normality are common in the adjusted distributions, potentially impacting the reliability of inferential tests that assume a symmetric distribution. An alternative smooth value function in place of the typical piece-wise function is explored to observe whether it yields desirable statistical properties for inferential purposes. The key contribution of this chapter is the finding that a skewed distribution is a likely consequence of applying risk transformations and that the practitioner should abstain from assuming symmetry when confidence intervals are computed for hypothesis testing. The BCa$^-$ bias-corrected accelerated bootstrap approach is proposed and evaluated against other candidate testing approaches when pairs of systems are the evaluation target. A greater understanding of the distributional properties of risk-adjusted score distributions at the paired-testing level helps to inform the practices required for computing risk inference over many systems (explored in subsequent chapters).

Chapter 3 addresses:

> **Research Question (RQ3)**: How do the distributional properties of risk-adjusted scores impact the results of parametric statistical tests?

1.1.2 Modeling Risk-Adjusted Scores On Many Systems

Chapter 4 commencing on page 95 expands on the outcomes of the previous chapter to inform a novel approach towards performing risk inference over multiple systems, where Bayesian inference is used with multiple comparison corrections for the first time in evaluating ad-hoc IR systems. Bayesian modeling is applied to pairs of systems at a time to validate the appropriateness of modeling the scores as skew-normal as opposed to a symmetric normal distribution. That skew-normal model is extended to model the effects of multiple risk-adjusted systems at once, along with topic effects, using a hierarchically modeled parameterization. The inferential outcomes of the proposed BRisk$^-$ method are compared against the paired frequentist alternatives with their appropriate corrections applied.

Chapter 4 responds to:

> **Research Question (RQ4)**: How can risk-adjusted scores be modeled over multiple systems with multiple comparison correction?

1.1.3 Bayesian Post-Hoc Risk Analysis On Many Systems

Chapter 5 starting on page 127 explores the merit of modeling traditional IR scores using Bayesian inference without risk, examining whether modeling the scores using distributions that more closely align with IR effectiveness data yield different outcomes. Participation in the TREC COVID evaluation exercise motivated the exploration, as the submitted run `RMITBFuseM2` moved up 35 system rankings on the RBP effectiveness metric when more complete judgments were available. This run rose the most places out of all submitted runs and received a special mention in Ellen Voorhees' SIGIR keynote, highlighting that interpreting score volatility and confidence is as important as ever when examining system dominance. The PPDRisk^{-} overlay is proposed, utilizing the posterior predictive distribution of simulated Bayesian models on traditional IR scores with its utility compared with the BRisk^{-} method proposed in the previous chapter. The approach is demonstrated to be more powerful than BRisk^{-}, yielding the first practical approach for performing risk inference over multiple systems that corrects for multiple comparisons, and honors the risk-adjusted score asymmetry identified in Chapter 3.

Chapter 5 answers:

> **Research Question (RQ5)**: How can standard IR scores be modeled over multiple systems (with multiple comparison correction) to improve the sensitivity of multiple system risk inference?

1.2 Organization

The thesis is organized in the typical structure, so as to guide the reader towards and through these three research areas. Chapter 2 on the following page surveys the history of empirical evaluation and experimentation in information retrieval, paying close attention to applying statistical inference and risk overlays on effectiveness metrics. Chapter 3 commencing on page 59 presents an analysis of the statistical properties of risk-adjusted score distributions, informing the novel approach taken in Chapter 4 starting on page 95 that directly models the risk-adjusted scores over many systems. Although modeling risk-adjusted scores directly is successfully achieved, the statistical power of the approach is insufficient for practical use. That leads into Chapter 5 commencing on page 127, which readdresses the above problem using the novel approach of using the posterior predictive distribution of standard IR scores with superior results. Finally, the contributions of the thesis are summarized in Chapter 6 on page 165 with future work.

2

Background

The contributions of this thesis draw from a broad range of topics, including *information retrieval* (IR) evaluation, frequentist and Bayesian statistics, and risk analysis. The chapter begins with an introduction to information retrieval evaluation and its origins in Section 2.1. Next, Section 2.2 on page 22 provides an overview of how these evaluation metrics tend to be computed and interpreted. Section 2.3 commencing on page 30 describes statistical testing in an IR context and how community consensus has evolved over the years. Section 2.4 on page 48 provides a brief overview of the statistical distributions of interest in this study. Section 2.5 starting on page 50 explores the various risk analysis techniques employed on IR systems and the results transferred from economics that inspired their combination with IR evaluation metrics. Section 2.6 on page 56 concludes the chapter with an exposition on the research gaps in the area of IR inferential risk analysis, which correspond to the research questions listed in Section 1.1 on page 8.

2.1 Information Retrieval Evaluation

The phrase "information retrieval" is said to have been first used in an uncirculated report by C. N. Mooers in 1951 [51]. To appreciate the significance of information retrieval evaluation in the present day, an examination of how information needs were resolved prior to computing is useful context to understand how IR evaluation approaches evolved.

2.1.1 From Libraries To Computers

Hierarchical Search. Sanderson and Croft [168] recount the history of innovations in the library sciences relevant to the field of information retrieval, honoring the critical role libraries played in knowledge sharing prior to the early 20th century. Although using a library to learn detailed information about a topic is still useful today, it has fallen out of favor as the primary means of accessing information. Starting from the catalog, searching for library resources involves physically traversing shelves of books containing many thousands of titles. Whether an information need can be resolved in a library depends on whether it has the relevant material on hand, whether the searcher can easily find that resource and is accessible;

0		**500**	**Natural Science**
10	Bibliography	510	Mathematics
20	Book Rarities	520	Astronomy
30	General Cyclopedias	530	Physics
40	Polygraphy	540	Chemistry
50	General Periodicals	550	Geology
60	General Societies	560	Paleontology
70		570	Biology
80		580	Botany
90		590	Zoology
100	**Philosophy**	**600**	**Useful Arts**
110	Metaphysics	610	Medicine
120		620	Engineering
130	Anthropology	630	Agriculture
140	Schools of Psychology	640	Domestic Economy
150	Mental Faculties	650	Communication and Commerce
160	Logic	660	Chemical Technology
170	Ethics	670	Manufacturers
180	Ancient Philosophies	680	Mechanic Trades
190	Modern Philosophies	690	Building
200	**Theology**	**700**	**Fine Arts**
210	Natural Theology	710	Landscape Gardening
220	Bible	720	Architecture
230	Doctrinal Theology	730	Sculpture
240	Practical and Devotional	740	Drawing and Design
250	Homiletical and Pastoral	750	Painting
260	Institutions and Missions	760	Engraving
270	Ecclesiastical History	770	Photography
280	Christian Sects	780	Music
290	Non-Christian Religions	790	Amusements
300	**Sociology**	**800**	**Literature**
310	Statistics	810	Treatises and Collections
320	Political Science	820	English
330	Political Economy	830	German
340	Law	840	French
350	Administration	850	Italian
360	Associations and Institutions	860	Spanish
370	Education	870	Latin
380	Commerce and Communication	880	Greek
390	Customs and Costumes	890	Other Languages
400	**Philology**	**900**	**History**
410	Comparative	910	Geography and Description
420	English	920	Biography
430	German	930	Ancient History
440	French	940	Modern Europe
450	Italian	950	Modern Asia
460	Spanish	960	Modern Africa
470	Latin	970	Modern North America
480	Greek	980	Modern South America
490	Other Languages	990	Modern Oceania and Polar Regions

Table 2.1: Initial Dewey Decimal Classification System specified by Dewey [63, p. 12]

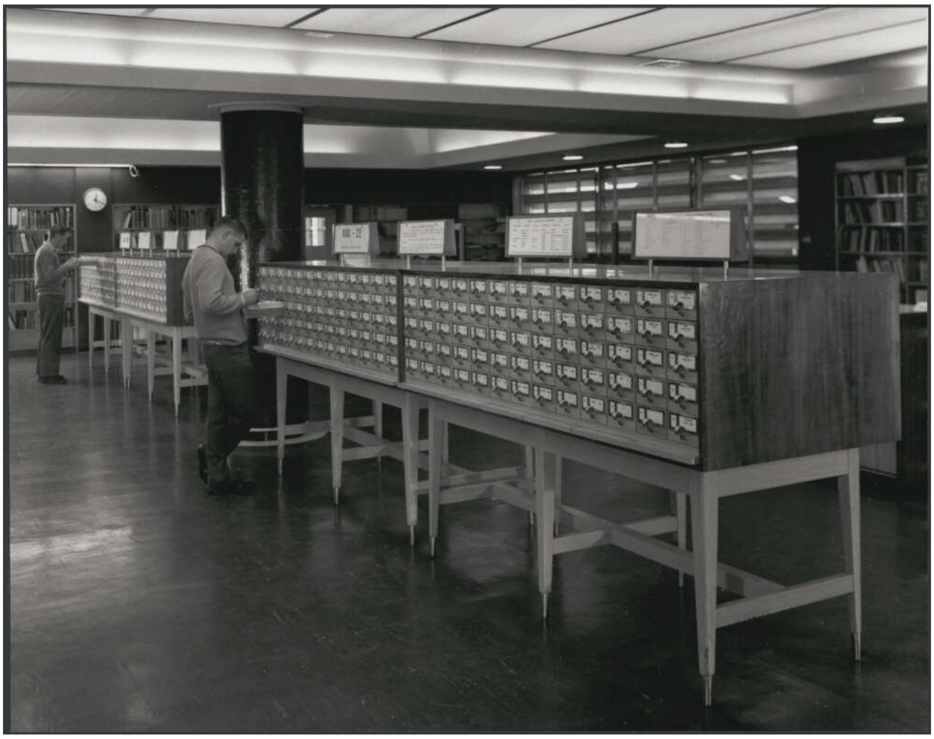


Figure 2.1: A photograph taken in 1961 of the card catalog used to enable efficient retrieval of physical resources in the Baillieu Library, University of Melbourne. [175]

both physically and in terms of the ability to comprehend the content. To ease the process of finding relevant library resources, libraries cluster alike resources on shelving, using systems such as the Dewey decimal classification system [63] presented in 1876. Table 2.1 on the previous page shows the initial set of topics established by Dewey [63]. Alternative classification systems exist across the world, such as the Colon system [172], and the Library of Congress classification system [4].

Keyword-Based Search. Despite the innovations in spatially organizing library resources, exploring whether a resource exists and narrowing down the scope of the physical search was a time-consuming task for resolving every information need. Figure 2.1 shows the card catalog used by patrons of the Baillieu Library at University of Melbourne, Australia, in the year 1961 to find resources. Rather than walking through the library and scanning the title of each book, library users were provided the alternative option of using card catalog to find their desired resource. There were multiple indexes within the card catalog; one for the author's name, another for the title of the resource, and a third for the subject of the resource. So each resource (a book, journal, newspaper, etc) may have multiple index entries (cards) in the catalog to make finding them easier. The key discovery from US government librarian Taube et al. [184] that resources can be effectively indexed by the subject headings was a major step forward in the evolution of information retrieval. That innovative index organization combined with mechanical and electronic inventions such as sorted punchcards, gave rise to the first incarnation of Boolean search [110]. Cleverdon [49] empirically validated the

	Relevant	Non-Relevant	
Retrieved	a	b	$a+b$
Not Retrieved	c	d	$c+d$
	$a+c=R$	$b+d=N$	

Table 2.2: The matrix adopted from Cleverdon [50, Figure 2] describing the statistics used to calculate the precision and recall measures.

effectiveness of the keyword-based search system by using a corpus of aeronautical manuals along with an established set of test queries. This major innovation transformed the keyword-based search task in the physical realm into a computer science problem, creating the first set-based automatic IR systems.

2.1.2 Important Early Innovations

The Cranfield Experiments. Cleverdon [50] is credited with proposing the "Cranfield" evaluation paradigm, which provided the first methodology for measuring the effectiveness of IR systems. In these experiments, $1{,}400$ documents of aeronautical manual data were used to measure methods of retrieving answers to a set of 225 queries. The relevance of each document against each query was determined by manual relevance assessment on a graded scale, with a score of 1 indicating "a complete answer to the question", and 4 being a reference of "minimal interest". Cuadra and Katter [60] highlighted issues with performing relevance assessments in general, noting that randomly selected experts often disagreed on the relevance of a document against a query, and individual experts disagreed with their own prior judgments. Despite those complaints, the Cranfield evaluation paradigm remains the default choice for evaluating IR systems. In the pioneering early days of IR evaluation the retrieved set of documents against queries were not ranked (much like the library card systems mentioned in the previous subsection). Due to that, set measures of effectiveness were used for quantitative evaluation. Table 2.2 describes the contingency matrix used to construct the definitions for *precision* and *recall*, which remain important concepts to this day. Precision is concerned with the ratio of relevant documents a retrieved against the total number of documents retrieved, $a+b$. Sharing a common numerator, recall is defined as a against the total relevant documents in the collection, $a+c$.

Cleverdon [50] observes that precision and recall tend to be inversely related. For example, when precision is high that often implies that recall is low, and when recall is high, precision tends to be low. Rather than reporting multiple values for precision and recall, Cleverdon [50] opted to plot the two quantities against each other on a single graph. When the amount of documents retrieved is varied, the resulting *recall-precision curve* develops an visual profile for the overall effectiveness of a system. Their experiments also varied the grading scale used to determine relevance, where relevant documents were graded within the ranges $\{[1],[1,2],[1,2,3],[1,2,3,4]\}$. Their analysis [50, Figure 10] favoured using the

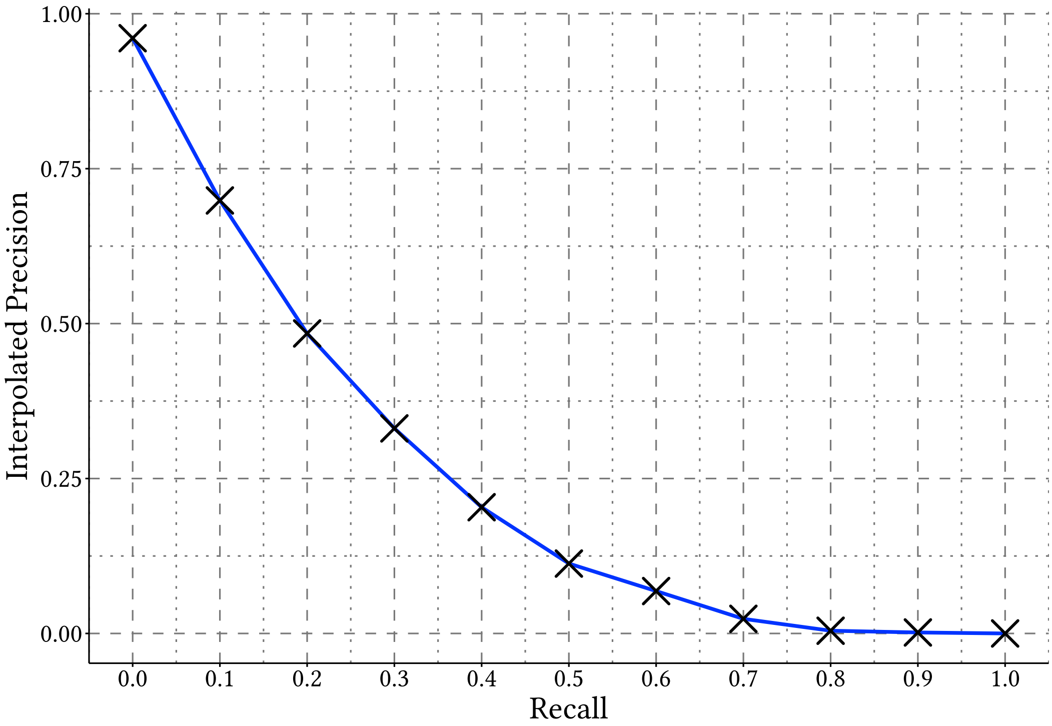


Figure 2.2: An example of a recall-precision curve.

binary classification range, {[1]}. Figure 2.2 provides an example of a recall-precision curve as an example. These recall-precision curves are still reported in TREC submission reports as supplementary data.

Usability. As new techniques for indexing and searching library resources emerged in the late 50s and early 60s, Mooers' law[1] [28, 138] highlighted the need for usable IR systems:

> "An IR system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for them not to have it."
>
> — Berul [28, p. 2–7]

Usability during this time was predominately focused on the user's ability to retrieve relevant documents using the system. As Frøkjær et al. [80] pointed out in more recent work, the usability of an IR system is interpreted by evaluating all three of effectiveness, efficiency, and satisfaction. *Effectiveness* is a measure of how well the system can retrieve relevant information for reliably resolving information needs, and is the central focus of this thesis. *Efficiency* relates to how quickly the system can perform the retrieval task initiated by the searcher. Lastly, *satisfaction* aggregates the attitudes of searchers when using the system, measured via questionnaires (Frøkjær et al. [80] recommends the SUMI inventory [112]).

[1] Quoted via Berul [28], who attributed the quote to the nonrecoverable resource Mooers [138]. The accuracy of the quoted text from the original source cannot be verified.

Ranked Retrieval. The introduction of ranked retrieval in the 1970s motivated a new era of IR evaluation. Jones [107] introduced the TF-IDF weighting scheme, which scales the importance of query terms (keywords) within documents towards resolving the information need of a searcher. The TF (term frequency) component counts the number of times a query keyword appears in a document. Further, IDF (inverse document frequency) is a measure of how rare the keyword is across the collection, computed by dividing the total number of documents in the corpus $|D|$ by the number of documents containing the keyword; logarithmically scaling the result. Multiplying TF and IDF together weighs the importance of a query term in a document, and summing the weight of each query term in a document yields the document *score*. Sorting documents by their score yields a *ranking* from a retrieval system.

Another important discovery in the 70s was relevance feedback [152]. In the Rocchio [152] approach, a human-in-the-loop would assess the relevance of a set of documents on an initial retrieval for a query. Documents marked by the searcher as relevant are used for positive re-enforcement of the query, and non-relevant documents are used to penalize terms appearing in those documents in the query. After that labelling process, a system-defined number of new terms are added to the query with their importance weightings varied by the user's feedback. The augmented query is re-issued to the retrieval system where the searcher is presented with a new ranked list of documents. Pseudo-relevance feedback is a fully automatic approach, which instead uses the top-ranked documents from the initial retrieval for positive re-enforcement of the user's query. That is, their query is augmented and reissued to the retrieval system unbeknownst to them, displaying one ranked list of documents to the searcher.

Salton et al. [166] encoded the TF-IDF weighting scheme into a vector space model for ranked retrieval of documents. The cosine similarity between a searcher's query vector and each document vector in the corpus is computed, to provide a real-valued similarity score for each document. Document across the collection are then ordered by their similarity score towards the user's query, providing a ranking. The Cranfield evaluation methodology was used in conjunction with recall-precision graphs to evaluate the vector space model, using aerodynamics, medicine, and world affairs document collections. Salton et al. [167] further applied the model for ranked retrieval of Boolean queries, as at this time, Boolean queries produced unordered sets of documents matching the terms in the query instead of ranked result lists. Biomedical, library science, electrical engineering, and computing corpora were used to evaluate the effectiveness of the *extended* Boolean retrieval model. When reflecting on the significance of the Cranfield evaluation paradigm, Cleverdon [51] highlighted ranked retrieval as being a significant achievement in the then quarter-century of its use in the discipline of information retrieval.

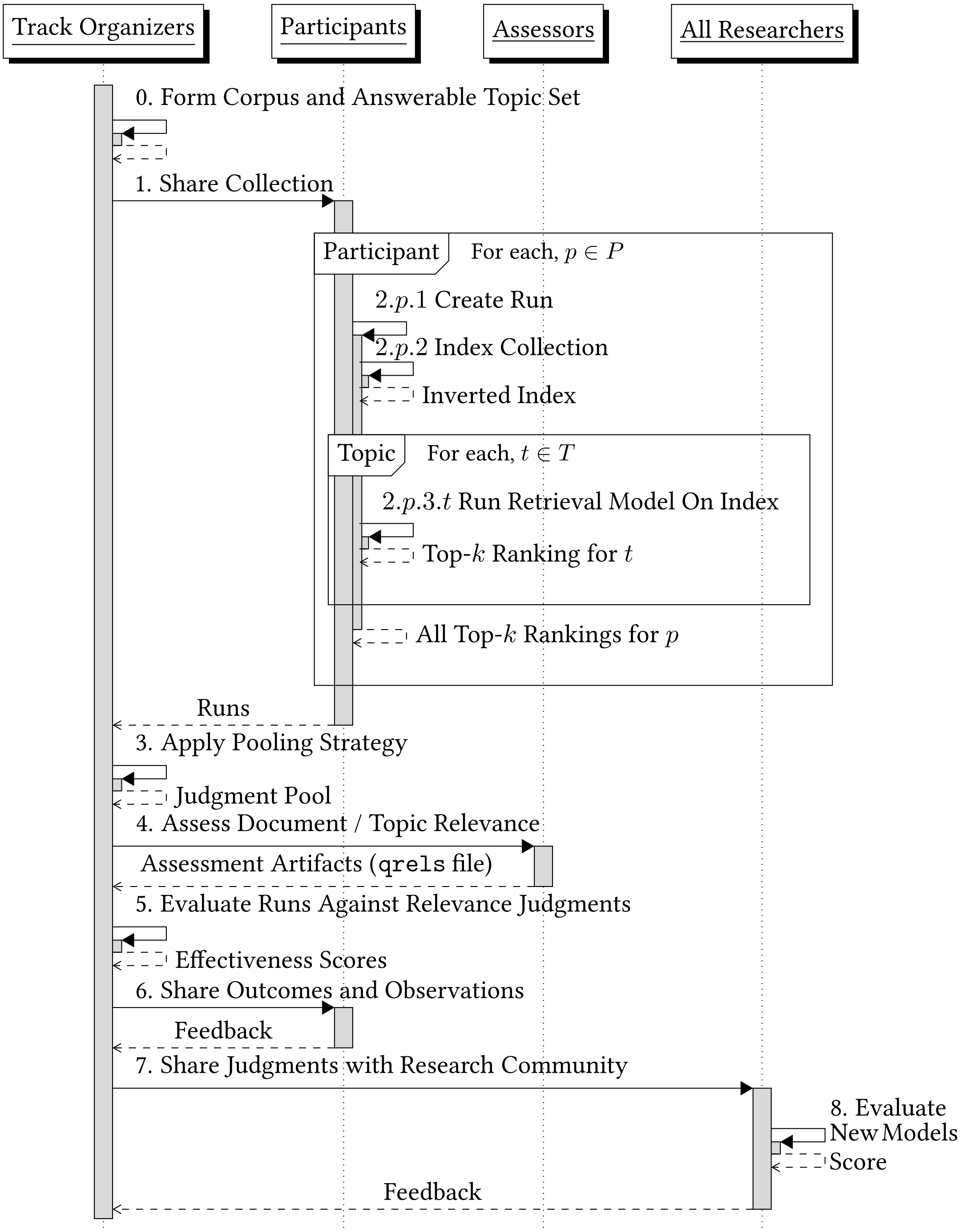

Track Organizers
Participants
Assessors
All Researchers
0. Form Corpus and Answerable Topic Set
1. Share Collection
Participant
For each, $p \in P$
2.p.1 Create Run
2.p.2 Index Collection
Inverted Index
Topic
For each, $t \in T$
2.p.3.t Run Retrieval Model On Index
Top-k Ranking for t
All Top-k Rankings for p
Runs
3. Apply Pooling Strategy
Judgment Pool
4. Assess Document / Topic Relevance
Assessment Artifacts (qrels file)
5. Evaluate Runs Against Relevance Judgments
Effectiveness Scores
6. Share Outcomes and Observations
Feedback
7. Share Judgments with Research Community
8. Evaluate New Models
Score
Feedback

Figure 2.3: A sequence diagram describing the steps involved in a pooled IR evaluation campaign, where the goal is to provide a set of reusable judgments that the research community can use to evaluate new models.

2.1.3 Pooled Offline Evaluation

Although the Cranfield methodology was (and remains to this day) the most widely adopted evaluation methodology for IR systems, there was an appetite for reform in the IR community. As Voorhees and Harman [204] mention in the history of TREC, in 1981 Jones [108] commented on the need for consolidation of the evaluation metrics and collections used. Further, it was unclear how Cranfield could be used for evaluating retrieval effectiveness on large corpora, due to the inability to exhaustively assess the entire collection for each query. The National Institute of Standards and Technology (NIST) was approached in 1990 by members of a US defense research project named TIPSTER, with the goal of understanding how to support search over a corpus containing a million documents. Voorhees and Harman [204] remarked that the 2 GB TIPSTER corpus was substantially larger than the closest comparable 2 MB newswire collection. Voorhees and Harman [204] note that the idea of *pooled evaluation* was first discussed by Jones and Rijsbergen [109], where a document collection too large to be completely assessed for topical relevance is filtered to an assessable size. The contributions of ranked document lists from many retrieval systems are used to improve the likelihood of finding relevant documents towards the topics. Pooled evaluation had not yet been fully explored until the TIPSTER initiative.

To evaluate how well retrieval models could work on a gigabyte-scale corpus, in 1992 NIST initiated the Text REtrieval Conference (TREC) to share the collection with the research community [94, 203]. Research groups indexed the collection and applied different retrieval models to explore their relative effectiveness, submitting their ranked result lists as *runs*. Figure 2.4 on the following page presents an example of the type of topics and document data supplied to participants in the first TREC ad-hoc retrieval task. As exhaustively judging the relevance of millions of documents against dozens of topics was an impossible undertaking, pooled evaluation was examined for the first time. The submitted runs were used to build a pool of documents to reduce the human-effort involved in assessing document relevance, where each run was evaluated using the same set of relevance judgments and evaluation metrics. Finally, the judgments were shared with the wider research community to evaluate runs that did not contribute to the judgment process as reusable *artifacts*. Zobel [224] found that evaluating unpooled retrieval systems using partially-pooled judgment sets produced consistent evaluation results, if a deep enough pooling depth was employed, with 100 per-topic suggested. Figure 2.3 on the previous page presents a sequence diagram illustrating the methodology used in a pooled IR evaluation campaign.

The ability to evaluate different retrieval models on a large collection using a consistent set of metrics was a major advance for the IR community. Buckley and Voorhees [35, p. 54] indicated that the previous reporting of evaluation metrics across papers had been inconsistent, with combinations of reporting "precision-at-10 (P@10), recall measures, utility, full recall-precision curves, three-point averages from the recall-precision curves, ten-point averages, and eleven-point averages." Immediately after the first TREC, the interpolated precision and recall curve evaluation techniques were superseded by the non-interpolated average precision approach, now known as AP [35] (covered in the evaluation metrics section in Section 2.3 on page 30). As AP is the most widely reported and used metric for IR evaluation of TREC ad-hoc test collections, the TREC committee achieved their goal of providing a standardized set of

TREC Topic

```
<top>
<head>  Tipster Topic Description
<num>  Number:  071
<dom>  Domain:  Military
<title>  Topic:  Border Incursions
<desc>  Description:
Document will report incursions by land, air, or water into the border area of one
country by military forces of a second country or a guerrilla group based in a second
country.
<smry> Summary:
Document will report brief incursions by land, air, or water into the border area of
one country by military forces of a second country or a guerrilla group based in a
second country.
<narr>  Narrative:
A relevant document will name the invading country or group, name the invaded country,
and identify the target or goal of the attacking force.  The target or goal must be a
military objective. It should NOT be about a war or invasion in which troops remain on
foreign soil for a sustained engagement.  It should NOT be about a territorial dispute
or economic dispute (such as fishing rights) which is being negotiated or argued by
civilian groups.
<con>  Concept(s):
1.  rebels, refugees, soldiers, guerrillas
2.  raid, assault, skirmish, dispute, fighting, clash, retaliation
3.  border, frontier, ford, thalweg, air space, territorial limit
4.  NOT sustained engagement
5.  NOT fishing rights
<fac>  Factor(s):
<def>  Definition(s):
</top>
```

Relevant Document

```
<DOC>
<DOCNO>AP890101-0013</DOCNO>
<HEAD>Israel Intensifies Search For Guerrilla Infiltrators</HEAD>
<HEAD>An AP Extra</HEAD>
...
<TEXT>
Israeli soldiers with powerful binoculars peer into the rocky expanse of southern
Lebanon after warnings that Palestinian guerrillas opposed to the PLO's peace overtures
may try to infiltrate Israel. ...
</TEXT>
</DOC>
```

Non-Relevant Document

```
<DOC>
<DOCNO>AP890101-0016</DOCNO>
<HEAD>Eritrean Rebels Reportedly Agree To Negotiate With Ethiopia</HEAD>
...
<TEXT>
An Eritrean rebel group that has waged Africa's longest civil war has agreed to
negotiate with the Ethiopian government for an end to the 26-year-old rebellion, a
Sudanese newspaper reported Sunday. ...
</TEXT>
</DOC>
```

Figure 2.4: One of many topics shared with TREC participants in the in the first ad-hoc retrieval task on the first volume of the TIPSTER corpus. Beneath the topic data are two 1989 *Associated Press* news articles in TRECTEXT format; one relevant and one non-relevant.

metrics that could be used to evaluate the effectiveness of IR models longitudinally. Armstrong et al. [15] reported AP to be the most widely used metric among papers published in top research venues between 1998–2009.

Another major contribution from the standardized evaluation approach was the observation that retrieval models vary in effectiveness over different topics and corpora. Multiple studies found that executing a search using many query variations and fusion techniques improved search effectiveness [20, 21]. Banks et al. [19] found that the topic effect is substantially larger factor than the system used to perform the search, a factor that has been exploited in recent studies [17, 18, 137]. TREC tracks led to further developments in IR ranking approaches, such as BM25 [151], language modelling [148], and learning-to-rank [120].

The scope of TREC has expanded to include web document collections (among many other retrieval tasks), with Hawking and Thistlewaite [97] establishing the groundwork with the TREC Very Large Collection (VLC) track. From the VLC beginnings came the Hawking and Craswell [96] overview of the first TREC Web track held in 2001, where many more recent tracks have been run since [48, 55, 56]. The success of pooled evaluation initiated several further activities with novel objectives, such as CLEF[2], NTCIR[3], and FIRE[4].

2.1.4 Evaluating Search By User Behaviors

User Studies On Offline Metrics. Table 2.3 on the following page presents a survey of the different types of collections and parameters used in the evaluation of offline IR systems involving a user study.

Hersh et al. [99] first posed the question of whether batch evaluation outcomes matched human interactions of the same systems, where a TF-IDF system is compared to an Okapi BM25 system. They carried out an instance recall study, where the number of saved relevant documents by a searcher is then compared against the number of known relevant documents. Although BM25 scored higher than TF-IDF in batch evaluation, there was no significant difference in instance recall task performance. Allan et al. [12] found that improvement in retrieval effectiveness resulted in greater user effectiveness, and that improvements in bpref from 50% to 60% provided the most benefit on hard topics, but not for easy topics. Turpin and Scholer [193] evaluated user performance using a precision-based task, finding that only whether the first document in a ranking was relevant was a weakly significant factor in determining whether a user was successful. Although Thomas and Hawking [185] did not compare two different ranking models, two of the same rankings with different document summaries were placed side-by-side, and searchers were asked to identify which of the two rankings was the most preferable. Thomas and Hawking were unable to identify a significant difference in preference for either query summarization approach. Al-Maskari et al. [10] found that success on the user task was predicted well by the offline evaluation method used, but only on the first

[2] CLEF homepage: `https://www.clef-initiative.eu/`, accessed on 21st July 2022.

[3] NTCIR homepage: `https://research.nii.ac.jp/ntcir/index-en.html`, accessed on 21st July 2022.

[4] FIRE homepage: `http://fire.irsi.res.in/fire/2022/home`, accessed on 21st July 2022.

	Year	Corpus	Topics	Workers	Systems	Top-k	$\setminus$	$\equiv$
Hersh et al. [99]	2000	TREC-8	6	24	2	50	✗	✗
Turpin and Hersh [191], [192]	2001	TREC 8/9	14	25	2	10	✗	✗
Allan et al. [12]	2005	Many News	45	33	1 [b]	50	✗	✗
Turpin and Scholer [193]	2006	WT10g	47 [a]	30	200	100	✗	✓ [d]
Thomas and Hawking [185]	2006	Google	306	23	1	20	✓	✓
Al-Maskari et al. [9]	2007	Google	104	26	1	10 [c]	✗	✓
Al-Maskari et al. [10], [7], [8]	2008	TREC-8	56	56	2	10	✗	✗
Smith and Kantor [176]	2008	Google	12	36	3	20	✗	✓
Zhu and Carterette [222]	2010	Yahoo!	30	25	1	10	✓	✓ [d]
Sanderson et al. [170]	2010	ClueWeb12B	30	296	19 [b]	10	✓	✓
Smucker and Jethani [178], [177]	2010	AQUAINT	8	48	1,000	10	✗	✓
Moffat et al. [136]	2013	Yahoo!	6	34	1	10	✗	✓
Sakai and Zeng [163], [164]	2019	NTCIR 9	100	15	15	10	✓	✗

[a] Topics were derived from real search sessions.
[b] Systems were synthetic run generations derived from relevance judgments.
[c] Result list size was variable and not always k, unlike the other studies listed.
[d] Query-biased summary presentations did not embolden the search keywords.

Table 2.3: A survey of publications where different systems or presentations on offline datasets were compared by presenting their outputs to human assessors. The $\setminus$ column indicates whether a side-by-side presentation of SERPs occurred, and the $\equiv$ column shows whether the SERPs included a query-biased summary. Notes beneath the table describe experimental differences and commonalities between studies.

few rank positions when evaluating interactively. The key observation from these studies (and other studies surveyed without a direct mention) are that they are not as sensitive in recognizing differences in retrieval effectiveness as offline batch evaluation approaches are.

Most recently, Sakai and Zeng [163] found that popular offline evaluation metrics such as NDCG correlate well with the ranking preferences of searchers. In those experiments, Sakai and Zeng hired 15 assessors to judge the SERPs of 15 systems submitted to NTCIR-9, a test collection with 100 topics. All 10,500 triplets of topic, s_1 ranking, and s_2 ranking were created, and then filtered to 1,127 combinations to ensure the ranking pairs were sufficiently different from each other with appropriate metadata. Of the SERPs presented, assessors selected whether they preferred the left, right, or neither presentation, using a forced choice approach. The study received better agreement for document ranking preferences between users than what was expected from previous user studies.

Evaluating Web Search. About the time where web search had become the most renown instance of IR, Broder [33] proposed a taxonomy of the topics that are searched for most frequently, which are still relevant today. The three main categories of web IR topics are:

- *navigational*: where the intent of the searcher is to access a website.
- *transactional*: the searcher wishes to immediately perform a transaction, such as downloading software or media.
- *informational*: the user is expected to read many documents to understand a topic.

where the type of evaluation metric used can be tailored to the topic's classification. For example, when a user issues a navigational query to a search engine, it is posed that one document satisfies this information need, and should be returned as close to the head of the result list as possible. The reciprocal rank metric (discussed in Chapter 1 and formalized in Section 2.2.1) captures this goal. Evaluating a dynamic document collection such as the web has more challenges than static collections, particularly around the issue of the *freshness* of judgments.

In offline evaluation, relevance assessors assign a judgment of topical relevance to a document suggested by one or more contributing systems as being relevant to that query. Rather than labeling a web-page by its *aboutness*, Jiang et al. [105] propose a relevance measure that models a user's utility of seeing a page in a ranking against the topic. But, the Internet is a corpus that is larger than normal offline collections and an assessment of relevance to a query may become stale over time. Yi et al. [217] state that in order to deal with that problem of scale, different metrics are used by large commercial search engines. Thomas et al. [186] show that effectiveness metrics for commercial search engines not only evaluate document rankings, they also consider factoid answers presented, advertisements, and aggregated information.

Hofmann et al. [102, Section 3] survey the literature in their "Metrics for Online Evaluation" section. They identified that the most common observations for evaluating result list effectiveness were to compare the rank of the links that were most often clicked [31], the time taken to click a webpage [78], and abandonment. They explain that in order for a fair comparison to be made, the above absolute measures can be compared within the same time interval, or an interleaving experiment could be used. Interleaving involves a *mixing rule* for balanced insertion into a result set to present to a user, and a *scoring rule* based on the observations that show users prefer a particular system. The interested reader is referred to Hofmann et al. [102] for a detailed review of online measures of IR effectiveness.

2.2 Effectiveness Measures

Effectiveness measures quantify an IR system's ability to retrieve relevant documents from a corpus in response to a query. The relevance assessments are made by human annotators, as discussed in Section 2.1. (A concrete example scenario where a practitioner might opt to use evaluation measures was presented in Figure 1.2 on page 5.) Many measures have been proposed over the years, evolving from the original *precision* and *recall* metrics explained in Section 2.1. Lu [122, Figure 2.7] identified 27 measures in their PhD thesis. This section provides an overview of the most common measures used in IR evaluation. Lu et al. [123] distinguished IR measures as either utility-based and recall-based. Utility-based metrics quantify the relevant documents retrieved by the system visible to the searcher. In contrast, recall-based measures quantify the effectiveness of a retrieved document list in terms of the number of relevant documents identified in the collection.

Preliminaries. Researchers in the IR community regularly wish to evaluate the effectiveness of a ranked list of documents quantitatively. Provided that a set of relevance judgments exists for a topic represented by a ranked list of documents on the same corpus, the ranking can scored using effectiveness measures.

A ranked list is an ordered sequence of document identifiers. Each document identifier in that list is looked up in the relevance judgments from rank $i = 1$ to the maximum number of documents retrieved k, constructing a relevance vector r of $r_i \in [0.0, 1.0] \cup \{?\}$ fractional values. That is, allow a $r_i = 0.0$ to indicate that the ranked document at i is non-relevant, $r_i = 1.0$ maximally relevant (bound by the maximum relevance value in the judgment set), and $r_i = ?$ when d_i is not judged on the topic being evaluated. Some measures treat r as $\lceil r \rceil$, where documents are interpreted as either entirely relevant or non-relevant.

The effectiveness metrics described in this section are computed against r for a system ranking on a given topic. When evaluating the effectiveness of a retrieval system, each query for every topic is submitted to the ranker to generate many rankings, where they are independently scored using the same metric. The arithmetic mean of these metric outputs summarize the system's effectiveness. The geometric mean has also been considered as an alternative to the arithmetic mean, noting that the geometric mean could be used for any measure with non-zero outputs described in this section to summarize system effectiveness. These statistics are further discussed when the *average precision* (AP) measure is defined in Section 2.2.2.

Precision and Recall. The precision and recall measures calculate ratios of the number of relevant documents in a set irrespective of their ranking. Both measures are often reported for a given rank cut-off, where unjudged documents are treated as non-relevant. The precision and recall metrics are defined as:

$$\mathit{Prec@k} = \frac{\sum_{i=1}^{k} \lceil r_i \rceil}{k} \ ; \qquad (2.1) \qquad \mathit{Recall@k} = \frac{\sum_{i=1}^{k} \lceil r_i \rceil}{R} \ , \qquad (2.2)$$

where k is the depth of the list to evaluate up to, and the total number of relevant documents for the topic under evaluation is R. Consider the relevance vector:

$$r = \langle 1.00, ?, 0.33, 0.00, 0.00, ?, 0.66, 0.00, 0.33, 0.00, 0.00 \rangle \ , \qquad (2.3)$$

and suppose that R is known to be 10. The precision and recall when $k = 5$ uses the shared numerator $\sum_{i=1}^{k} \lceil r_i \rceil = 2$. Therefore, *Prec*@$5 = 2/5$ and *Recall*@$5 = 2/10$. The *Prec*@k metric is mentioned with the shorthand form P@k in this thesis and throughout the literature. In addition, ranked lists are frequently truncated to various sizes depending on the evaluation goals of their respective studies, where k values of: 5, 10, 20, 50, 100, and 1000 are common.

2.2.1 Utility-Based Measures

The class of evaluation metrics that measure the usefulness of a ranking without considering whether documents known to be relevant could have been retrieved is referred to as utility-based. For example, P@k is a utility-based metric, as it only evaluates the quality of a list with respect to the utility a searcher can derive. However, P@k is a set-based utility metric, meaning that it does not consider the order of the documents in the ranked list. As ranked retrieval is the most common form of information retrieval, most utility-based metrics measure the gain of a document as a function of its position in a SERP and relevance. In the most simple case, the reciprocal rank of the first relevant document in a ranking can be used to measure the effectiveness of a ranking.

Reciprocal Rank. *Reciprocal rank* (RR) is a utility-based measure that is typically used to evaluate the effectiveness of navigational topics on a search engine. It is defined as the reciprocal of the rank of the first relevant document in a ranked list. If there are no relevant documents in the retrieved set, then the RR score is 0. The RR metric models a searcher who will stop looking for relevant documents once they find one, a natural fit for evaluating the effectiveness of queries to find a website [33]. The utility a searcher receives from finding the first relevant document is the same as the RR score. Since the relevance grade of a document is not considered, judged r_i values may be considered binarized. The RR measure is defined as:

$$\mathit{RR} = \begin{cases} 0 & \text{no relevant documents retrieved,} \\ 1/i & \text{where } \lceil r_i \rceil \text{ is the first relevant document in } r. \end{cases} \tag{2.4}$$

There are, however, cases where the relevance grade of a document should be considered, particularly for informational topics where the searcher may need to read many documents to find the one that answers their question. A ranking with a high density of relevant documents at the head of the list is useful, as the searcher does not need to scroll through many non-relevant documents to find their required information. Metrics that quantify this iterate over r and compute the weighting of documents at each position i in the ranked list, then sum the gains for a total utility score. Weighted precision metrics Lu [122, Equation 2.20] adopt the form:

$$\sum_{i=1}^{|r|} w(i) \cdot g(r_i)\,, \tag{2.5}$$

where $w(i)$ reweighs the importance of relevant documents as a function of their rank in the list, and $g(r_i)$ is a gain function that transforms relevance values to suit the needs of the metric. This subsection now considers weighted utility-based metrics of the Equation (2.5) form.

Rank-Biased Precision. *Rank-biased precision* (RBP) by Moffat and Zobel [135] is a metric with a variable *persistence* parameter ϕ that controls the emphasis placed on quantifying the density of relevant documents present at the head of a ranked list. The metric models a searcher who does not wish to continuously examine each document in a ranked list. They will look at the first document in a list and examine the next one with probability ϕ, or stop looking with probability $1 - \phi$. If the searcher does examine the following document, the process is repeated, independently of whether they arrived at a relevant document, until the searcher stops looking. The RBP weighting function for a document at rank i is therefore defined as:

$$w_{\phi}(i) = \phi^{i-1}(1 - \phi)\,, \tag{2.6}$$

where the expected viewing depth d of a searcher in a session can be used to estimate ϕ, using the formula:

$$\phi = \frac{d-1}{d}\,. \tag{2.7}$$

The equation to compute the typical RBP observed score for a r where only relevant documents will make a score contribution is:

$$\mathit{RBP} = \sum_{i=1}^{\infty} w_{\phi}(i) \cdot g(r_i)\,, \tag{2.8}$$

where $g(r_i)$ is the identity function.

In addition to reporting the RBP *minimum* score, it is also possible to compute a score *residual*, which helps practitioners to understand the extent to which the observed score is due to unjudged documents in the ranked list. The observable *maximum* possible RBP of the retrieved list is also computed using Equation (2.8), where unjudged documents are treated as fully relevant. The RBP residual is then the maximum score less the minimum score. Due to the user model used to compute the RBP weighting function, the RBP residual includes the utility of potentially finding relevant documents outside of the system-based decision to retrieve $|r|$ documents. To calculate the infinite component of the RBP residual, the observed RBP residual is supplemented with the additional score uncertainty supposing every document retrieved by the system beyond the rank cut-off of $|r|$ would be maximally relevant.

Therefore, the complete RBP residual is the observed list residual summed with the infinite component:

$$(1-\phi) \sum_{i=|r|+1}^{\infty} \phi^{i-1} = \phi^{|r|}\,;\; \mathit{RBP}_{\mathit{resid}} = [\mathit{RBP}_{\mathit{max}} - \mathit{RBP}_{\mathit{min}}] + \phi^{|r|}\,. \tag{2.9}$$

A high RBP residual indicates that the observed score may be higher if the unjudged documents were considered relevant. Analyzing the residual value may guide the choice of other reported metrics and the ϕ parameter to ensure the evaluation goals are compatible with the constraints of a given collection, where a lower ϕ value provides more score certainty.

Expected Reciprocal Rank. The *expected reciprocal rank* (ERR) metric by Chapelle et al. [45] is a weighted precision metric that models the probability that a user will stop examining documents after finding a relevant document that satisfies their information need. To compute the likelihood in an offline setting, Chapelle et al. [45] transform the relevance grades into a probability distribution by normalizing the relevance grades between zero and one with the gain function:

$$g(r_i) = \frac{2^{r_i} - 1}{2}\,. \tag{2.10}$$

With that transformation to the relevance component, the ERR weighting component has the equation:

$$w(i) = \frac{1}{i} \cdot \prod_{j=1}^{i-1} 1 - g(r_j)\,. \tag{2.11}$$

Chapelle et al. note that when a real search engine is being evaluated, the g probabilities can be derived from click-through data. They also note that when $r' = \lceil r \rceil$ is provided and $g(r_i) = r_i$, the ERR metric produces equivalent results to the reciprocal rank metric.

Similar to the other metrics considered in this section, only relevant documents will contribute to the ERR observed score at each position in the ranked list. The ERR metric is defined as:

$$\mathit{ERR} = \sum_{i=1}^{|r|} w(i) \cdot g(r_i)\,, \tag{2.12}$$

where the product component is used to model the probability that a searcher will stop viewing the ranked list after viewing the ith document by stepping through each document's predecessor and summing their stopping probabilities.

Discounted Cumulative Gain. *Discounted cumulative gain* (DCG) is another utility-based measure used to evaluate the effectiveness of a ranked list. The DCG metric shares a similar user model with the RBP metric, where the user derives less utility from a document that is further down the ranked list; and will stop examining the ranked list independent of how many relevant documents they have viewed in the list. The CG component models a user that will examine every document in a ranking, defined as:

$$\mathit{CG} = \sum_{i=1}^{|r|} w(i) \cdot g(r_i)\,, \tag{2.13}$$

		RBP $\phi = 0.8$				ERR				DCG			
		min		*max*		*min*		*max*		*min*		*max*	
i	r	w	$w \cdot g$	w	$w \cdot g$	w	$w \cdot g$	w	$w \cdot g$	w	$w \cdot g$	w	$w \cdot g$
1	1.00	0.200	0.200	0.200	0.200	1.000	0.500	1.000	0.500	1.000	1.000	1.000	1.000
2	?	0.160	0.000	0.160	0.160	0.250	0.000	0.250	0.125	0.631	0.000	0.631	0.631
3	0.33	0.128	0.042	0.128	0.042	0.167	0.021	0.083	0.011	0.500	0.165	0.500	0.165
4	0.00	0.102	0.000	0.102	0.000	0.108	0.000	0.054	0.000	0.431	0.000	0.431	0.000
5	?	0.819	0.000	0.819	0.080	0.087	0.000	0.044	0.021	0.387	0.000	0.387	0.386
6	0.66	0.065	0.043	0.065	0.043	0.072	0.021	0.018	0.005	0.356	0.235	0.356	0.235
7	0.00	0.052	0.000	0.052	0.000	0.044	0.000	0.011	0.000	0.333	0.000	0.333	0.000
8	0.33	0.041	0.013	0.041	0.013	0.038	0.004	0.010	0.001	0.315	0.104	0.315	0.104
9	0.00	0.033	0.000	0.033	0.000	0.030	0.000	0.007	0.000	0.301	0.000	0.301	0.000
10	0.00	0.027	0.000	0.027	0.000	0.027	0.000	0.007	0.000	0.289	0.000	0.289	0.000
Score:		Σ	0.299	Σ	0.540	Σ	0.547	Σ	0.664	Σ	1.504	Σ	2.521
Residual:		$0.540 - 0.299$			0.241	$0.664 - 0.547$			0.117	$2.521 - 1.504$			1.017
$+\infty$:		$\phi^{10} = 0.108$			0.349				$\varnothing$				$\varnothing$

Table 2.4: Worked example evaluating RBP, DCG, and ERR, with their corresponding residuals. Only relevant documents are able to contribute a score to the metrics, and only the unjudged documents which could have been fully relevant are considered when calculating each residual. Note that RBP is a metric that is designed to be used with infinite lists, and therefore the score uncertainty of the retrieved list is combined with the possibility that unjudged relevant documents could have been provided by the system beyond the $|r|$ documents.

where $w(i) = 1$ and $g(i)$ is the identity function (g remains the identity function for each iteration on CG). The weighting component in DCG that discounts the gain of a document at rank i is defined by:

$$w(i) = \frac{1}{\log_2(i+1)}\,, \tag{2.14}$$

which is supplied to the DCG definition, denoted:

$$DCG = \sum_{i=1}^{|r|} w(i) \cdot g(r_i)\,. \tag{2.15}$$

After discounting, the DCG measure models a user that will find relevant documents at the top of the ranking more valuable than those found deeper. A DCG score exists in the range $[0, \infty)$, where the score is usually *normalized* (NDCG) by the maximum possible DCG score for a given list, a recall-based metric discussed in the next subsection.

Weighted Utility Example. Table 2.4 on the previous page provides an example of calculating the RBP, DCG, and ERR metrics with their corresponding residuals. The example is based on the same ranking of 10 documents as in Equation (2.3) on page 23. The RBP persistence parameter is set to $\phi = 0.8$, modeling a searcher expected to view the top-5 documents in a ranking. Although it is not a common practice to report the residual values on DCG and ERR, they could be used to provide additional insight into the caveats of reported scores. For example, the DCG and ERR residuals indicate that the observed score may be higher if the unjudged documents were judged as fully relevant. The RBP residual combines the uncertainty of the retrieved documents that were not judged with the potential for everything after the rank cut-off to be maximally relevant. For the given relevance vector r, the 0.299 RBP score seems benign. However, when the residual is considered, the uncertainty in the score is more considerable than the observed score. The score could have been as high as 0.648 if every unjudged document was fully relevant (including the infinite residual component). Large score residuals indicate that the completeness of the judgments available is at odds with the evaluation goals of the task. When the evaluation goal is altered to model $\phi = 0.5$ for a searcher expected to view the first two documents, the score is $\langle 0.523, +0.282 \rangle$, which is similar to the ERR score pair $\langle 0.547, +0.3 \rangle$.

2.2.2 Recall-Based Measures

Average Precision. *Average precision* (AP) is one of the most widely reported effectiveness measures in IR evaluation [15]. It takes the mean of the precision values at each relevant document in a ranked list, binarizing the relevance values and treating unjudged documents as non-relevant. Dupret and Piwowarski [68] show that the AP metric can be modified to consider graded relevance values; however, the modified metric has not been widely adopted. The AP metric has the definition:

$$AP = \frac{1}{R} \sum_{i=1}^{|r|} r_i \cdot P@i \,, \tag{2.16}$$

where P@i is the shorthand for *Prec@*i defined in Equation (2.1) and R is the known count of relevant documents in the collection for the topic under evaluation. Dupret and Piwowarski [68] note that the AP measure models a user who requires R-relevant documents to satisfy their information need, aiming to sequentially read each document in a ranking according to their predicted relevancy until R documents have been found.

In the Voorhees [200] TREC report for the Robust 2004 track, the track organizers identified the need to report an effectiveness metric that provides more emphasis to topics that perform poorly. Instead of aggregating the AP scores across runs using the arithmetic mean, the organizers proposed to use the geometric mean. The geometric mean of IR effectiveness scores for a system s is calculated by taking the product of each topic T score, and then taking the nth root of the product, where $|T|$ is the number of effectiveness scores calculated in s:

$$GM(s) = \left(\prod_{j=1}^{|T|} s_j \right)^{\frac{1}{|T|}} , \tag{2.17}$$

where every s effectiveness value is greater than zero. Maligranda [125] notes that the geometric mean is always lower than the arithmetic means due to the AM–GM inequality:

$$\frac{1}{|T|} \sum_{j=1}^{|T|} s_j \geq \left(\prod_{i=j}^{|T|} s_j \right)^{\frac{1}{|T|}} . \tag{2.18}$$

The *stability* of an evaluation metric can be explored when comparing their performance across different collections. Webber et al. [211] note that many methods exist for measuring metric stability. Most methods generate subsets of topics from a larger set and then analyze how consistently the aggregated measure determines which of a pair of systems was the more effective across topic sets. Webber et al. report that some methods for determining metric consistency only count statistically significant differences between pairs of systems for varied topic samples in their tallies. (Statistical significance testing is covered in the next section of this chapter). Voorhees [200] used the Buckley and Voorhees [34] approach to compare the relative stability of using the geometric mean instead of the arithmetic mean on the AP measure. Each relative score difference was tallied into counts of: better, worse, or the same (within a 5% threshold) when different topics were supplied. Buckley and Voorhees found that the geometric mean approach is less stable than the typical arithmetic AP calculation when 50 topic are evaluated, where the geometric measure has an error rate of 5.2% and the arithmetic statistic is 2.4%. For the geometric method to achieve a similar level of stability to the arithmetic mean, Buckley and Voorhees demonstrate that the geometric statistic requires 100 topics, which is ordinarily larger than the number of topics available in a test collection.

Normalized Discounted Cumulative Gain. The *normalized discounted cumulative gain* (NDCG) measure generalizes the utility-based discounted cumulative gain measure to bring the score into the range $[0, 1]$. The restricted range is useful for comparing the effectiveness of different topics on the same system. Easy topics with many relevant documents will dominate an averaged DCG score, making it difficult to compare the effectiveness of topics with fewer relevant documents. To fairly compare DCG scores across a collection, the DCG score is treated as a ratio of the *ideal DCG score* (IDCG) for the topic under evaluation:

$$NDCG = \frac{DCG}{IDCG} , \tag{2.19}$$

where DCG is computed using Equation (2.15). The IDCG value is computed by creating a synthetic document ordering based on the relevant documents for the topic under evaluation and then sorting them by their relevance value in descending order. Lu et al. [123] note that due to the IDCG normalization, a residual score component cannot be evaluated, which they note is the case for all recall-based measures.

Binary Preference. A recall-based measure that excludes unjudged documents from a ranking is binary preference (bpref). Similar to R being the count of relevant documents in the collection for the topic under evaluation, let N be the count of non-relevant documents. Allow that $r' = r \setminus \{?\}$, enabling the transformed relevance vector to compute the bpref measure:

$$\textit{bpref} = \frac{1}{R} \sum_{i=1}^{|r'|} \left[\lceil r'_i \rceil \cdot \left(1 - \frac{\min \left\{ \sum\limits_{j=1}^{i-1} 1 - \lceil r'_j \rceil, R \right\}}{\min\{N, R\}} \right) \right] , \tag{2.20}$$

where $\sum_{j=1}^{i-1} 1 - \lceil r'_j \rceil$ is the number of non-relevant documents above the current position i in the list during the summation. Moffat and Zobel [135, p. 12] note that the bpref measure does not have an obvious user model.

2.3 Statistical Inference

The previous section discussed how IR evaluation metrics quantify the effectiveness of ranked lists of documents. These metrics are primarily used to determine whether a system is more effective than one or many alternative systems. As the total number of submittable queries to any given search engine is infinite, and the cost of curating judgments for many thousands of topics is prohibitive, the evaluation of IR systems is generally performed on a subset of at least 50 topics. With this in mind, practitioners ordinarily use statistical testing to mitigate the risk of sample statistics being biased by the selected topic set when attempting to make broader claims about the effectiveness of a system.

Preliminaries. When assessing whether one or more sets of measured data values are *significantly* different from each other, significant means that the associated difference is unlikely to have occurred by chance. Concretely, suppose a pair of systems evaluated against 50 topics have mean effectiveness scores of 0.3 and 0.4 across topics, respectively. The practitioner wishes to test whether that difference in mean score can be attributed to chance; testing whether the *null hypothesis* holds. If the chance of the null hypothesis holding is less than 5%, then the researcher can be 95% confident in rejecting the null hypothesis, treating the differences in values as significant. Note that statistical tests are probabilistic and do not offer certainties. Two main types of errors can occur when performing null hypothesis statistical tests: a type I error happens when the null hypothesis is rejected when it should not have been, and conversely, a type II error occurs when the null hypothesis is accepted when it should have been rejected. Urbano et al. [197] discuss a third type III error, where the null hypothesis is correctly rejected for the wrong reasons.

There are two key philosophies for measuring the probability of an event occurring in statistics. The first and most commonly used approach in IR is frequentist reasoning, where confidence is derived from counting the occurrences of an event happening many times. Frequentist inferential methods yield *confidence intervals* (CIs) that bound a fixed population parameter derived from the sample observations. The second is the Bayesian approach, where the probability of an event occurring is tethered to prior beliefs in the event occurring. Bayesian *credible intervals* are interpreted against the prior distribution specified for the population parameter, treated as a random variable. Irrespective of the inferential paradigm, statistical tests also differ in their assumptions about the data and error characteristics. A statistical test which assumes the data follows a given distribution is said to be *parametric*, whereas a test which makes no such assumptions is *nonparametric*. Kitchen [113] comments that parametric tests are slightly more powerful when their assumptions have been met, but nonparametric tests are substantially more powerful when normality is violated.

This section first discusses the standard IR statistical tests in Section 2.3.1, which are used to compare pairs of systems at a time. Next, Section 2.3.2 on page 37 examines the techniques used in the IR literature to estimate how likely a statistical test is to detect a difference between two systems. Section 2.3.3 starting on page 39 explores how the paired testing techniques have been extended to compare multiple systems at a time. When comparing multiple systems, the number of hypotheses being tested increases the chance of false positives (a type I error). Section 2.3.4 on page 42 concludes the section with a discussion of Bayesian inference and how it has been used in IR.

2.3.1 Standard Statistical Tests

This section introduces the standard statistical tests used in the analysis of IR effectiveness data, all of which are frequentist. Each test is capable of detecting a significant difference between two groups of data, where the probability of each test is codified into a p-value in the range $[0.0, 1.0]$. A p-value of zero indicates no chance of an event occurring, and one means the opposite. Note that the common interpretations of p-values only apply in the context of frequentist testing. Instead of only reporting the p-value in statistical analysis, Sakai [159] has recommended also reporting the effect size (the minimum detectable score difference in units of standard deviations) of the test statistic, as sample size can play a more domineering role as to why the null hypothesis has been rejected.

Sign Test. The sign test is a nonparametric null hypothesis statistical test that counts the number of times a difference of paired values is positive or negative, which is used to estimate the chance that the two sets have been drawn from the same distribution. It has been used as a baseline in empirical IR to compare the statistical power of the tests presented in this subsection [145, 179, 197]. If the null hypothesis is true where there is no difference between the two lists, then the number of positive and negative differences should be approximately equal, where the tolerance of departure from equality is governed by the sample size of each set. Pairs of values with a difference of zero are removed from the calculation.

As only positive or negative differences in values between sets are considered useful, and the null hypothesis implies that a positive score will occur at the same rate as a negative score, the null can be modeled using a Bernoulli distribution. The Bernoulli distribution is most easily thought of as being used to calculate the probability of seeing x successful coin flips over n trials. As a fair test of the null hypothesis treats either side of the coin with an equal chance, it has the probability mass function:

$$f(x,n) = \binom{n}{x} 0.5^x \cdot 0.5^{n-x} = \binom{n}{x} 0.5^n \,, \tag{2.21}$$

where 0.5 gives each sign a half-chance of showing over many trials, and n is the sample size minus the count of differences in values which were exactly zero.

The minimum of the positive signs and the negative sign count is identified to calculate the p-value of the sign test; let this be known as *min*. Then, the sum of the probabilities of each possible value up to *min* yields the p-value of a *one-sided* test:

$$p = \sum_{i=0}^{\mathit{min}} f(i,n) \,. \tag{2.22}$$

When a null hypothesis test is one-sided, the researcher is testing with the expectation that one list yields larger or smaller values than the other. However, with IR score differences, an appropriate null hypothesis test should presume that both lists are drawn from the same distribution where there is no discernable difference between lists. Consequently, the probabilities of the other side of the distribution need to be accounted for. As fair Bernoulli distributions are symmetrical, the two-sided p-value is computed by $2p$.

Table 2.5 on the following page shows a worked example of evaluating the sign test on two pairs of hypothetical system effectiveness score sets. Across each column is a rank-relevance list in the same format discussed in the previous section. In the presented example, fifteen topics are being evaluated. Of the 13 non-zero differences in their scores, the Δ column shows the sign with respect to the difference (the 9th and 12th topic score differences are ignored). Note that the sign test does not account for the magnitude of the score difference. When tallying the negative and positive signs, the positive sign count is 3, forming the required *min* value. The probability mass function f is then computed over each step up to *min* and summed to get the one-sided p-value and doubled to yield the two-sided version IR practitioners use. As the two-sided p-value exceeds the traditional 0.05 value for 95% confidence, the sign test is not powerful enough to reject the null hypothesis on these observations.

	Topics														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
s_1	0.4	0.5	0.1	0.2	0.2	0.2	0.0	0.4	0.3	0.0	0.5	0.1	0.4	0.0	0.1
s_2	0.1	0.1	0.7	0.8	0.6	0.6	0.7	0.3	0.3	0.8	0.7	0.1	0.5	0.1	0.8
Δ	$+0.3$	$+0.4$	-0.6	-0.6	-0.4	-0.4	-0.7	$+0.1$	0.0	-0.1	-0.2	0.0	-0.1	-0.1	-0.7

$\lvert\Delta^+\rvert$	$\lvert\Delta^-\rvert$	*min*	n	$f(0,13)$	$f(1,13)$	$f(2,13)$	$f(3,13)$	$p = \Sigma f$	$2p$
3	10	3	13	0.000	0.002	0.010	0.035	0.047	0.094

Table 2.5: A worked example of the sign test, where the effectiveness of two systems is being compared for a metric across 15 topics. Their paired score difference Δ is taken, which is used to derive the counts of positive and negative score differences used for the sign test. Where score differences are zero, they are ignored in the sign test calculation. The *min* and n columns are the minimum and sum of each positive and negative difference, supplied in steps to the binomial probability mass function f defined in Equation (2.21). Summing the probabilities up to the *min* value gives the one-sided p-value of the test. The two-sided p-value is obtained by doubling the one-sided p-value, exploiting the symmetry of a fair Bernoulli distribution. Since the two-sided p-value exceeds 0.05, the sign test cannot reject the null hypothesis on the observed data.

Wilcoxon Signed-Rank Test. The Wilcoxon signed-rank test is a two-tailed nonparametric null hypothesis test, where evidence of rejecting the null is derived from relative ordinal rank changes between lists. Smucker et al. [179] note that the Wilcoxon signed-rank test was created to address the shortcomings of the sign test, where the sign test does not account for the magnitude of the difference between the two lists. Importantly, however, Urbano et al. [198] note that the Wilcoxon test is not strictly free of assumptions, as it supposes that the distribution of the differences is symmetric about the median. To apply the Wilcoxon signed-rank test, let the paired differences in scores with zeroes removed be defined as Δ'. Then, the sign from each of the values in Δ' are removed and sorted in ascending order. The ranks of each sorted value are then computed. If there are ties in the ranking, then the mean rank for each tied value is used. Finally, the original signs that were stripped for the purposes of ranking each of the items in Δ' are brought back in to index the rank positions. The sum of the ranks of the positive score differences yields W^+, conversely, W^- for negative deltas. Finally, the $\min\{W^+, W^-\}$ is determined to be the test statistic, which is compared to the critical value for a given number of observations $|\Delta'|$ and the required significance level.

Analytically evaluating the Wilcoxon W statistic into a p-value is a complicated procedure and falls outside the scope of this thesis.[5] A by-hand method exists for performing the test using a lookup table [131]. When the W value of the test statistic is smaller than the critical value in the W-lookup table for a given number of paired observations and required significance level, then the null hypothesis can be rejected.

[5] The interested reader is refered to the R implementation of the Wilcoxon density function: `https://github.com/wch/r-source/blob/trunk/src/nmath/wilcox.c`.

	Topics														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
s_1	0.4	0.5	0.1	0.2	0.2	0.2	0.0	0.4	0.3	0.0	0.5	0.1	0.4	0.0	0.1
s_2	0.1	0.1	0.7	0.8	0.6	0.6	0.7	0.3	0.3	0.8	0.7	0.1	0.5	0.1	0.8
Δ	$+0.3$	$+0.4$	-0.6	-0.6	-0.4	-0.4	-0.7	$+0.1$	0.0	-0.8	-0.2	0.0	-0.1	-0.1	-0.7

Δ'	$+0.3$	$+0.4$	-0.6	-0.6	-0.4	-0.4	-0.7	$+0.1$	-0.8	-0.2	-0.1	-0.1	-0.7
$abs(\Delta')$	0.3	0.4	0.6	0.6	0.4	0.4	0.7	0.1	0.8	0.2	0.1	0.1	0.7
$sort(abs(\Delta'))$	0.1	0.1	0.1	0.2	0.3	0.4	0.4	0.4	0.6	0.6	0.7	0.7	0.8
$avgrank(abs(\Delta'))$	2	2	2	4	5	7	7	7	9.5	9.5	11.5	11.5	13
$original\_sign(abs(\Delta'))$	$-$	$-$	$+$	$-$	$+$	$-$	$-$	$+$	$-$	$-$	$-$	$-$	$-$

W^+	W^-
14	77

	Confidence Level					
n	80%	90%	95%	98%	99%	99.5%
5	2	0				
10	14	10	8	5	3	1
13	26	21	17	12	9	7
15	36	30	25	19	15	12
20	69	60	52	43	37	32
25	113	100	89	76	68	60
...	...	...	...	...	...	...

Table 2.6: An on-paper example that tests for the null hypothesis on paired samples of IR effectiveness scores using the Wilcoxon signed-rank test. The example uses the same system-topic score observations presented in Table 2.5. The zero score differences are removed from the Δ in the Δ' row. These values then have their sign removed in one row, are sorted in the subsequent row, and then their average rank is tabulated. The original sign is included in the bottom row of the top table; used to reference which rank values to sum for the positive and negative score differences. As the W^+ value has the smallest sum, it is highlighted, and each rank component is highlighted in the *avgrank* row that contributed to this sum. As there were 13 paired observations after removing zeros, and the required confidence level is 95%, the Wilcoxon lookup table indicates that the critical value is greater than the W^+ value. Therefore, the differences are statistically significant.

Table 2.6 presents a worked example for performing a null hypothesis test using the Wilcoxon signed-rank test, using the same observed data as the sign test example. The same observations are used to demonstrate that different tests have different characteristics. (Note that it is not a sound methodological practice to rotate through statistical tests until the desired outcome is shown, nor is it appropriate to cycle through other metrics until significance is achieved, this is known as the *look elsewhere effect* [98, 212].) As the sum of the highlighted $avgrank(abs(\Delta'))$ values corresponding to positive score differences is smaller than negative scores, W^+ is the value used by the test statistic. The *critical value* of W is found by looking up the required 95% confidence level and the count of non-zero score differences $n = |\Delta'|$ in the rows. As $14 < 17$, the lists are significantly different using the Wilcoxon test.

Student t-Test. The Student t-test is widely used across many academic fields for null hypothesis statistical testing. The test supposes that a test statistic can be described by a t-distribution under the null hypothesis, making it a parametric test. However, it is considered to be robust to deviations from normality [104, 158]. In IR, that test statistic is generally the arithmetic mean, as that is the predominant statistic used to measure the relative difference in average effectiveness between systems. As the population variance of a sample is unknown (if complete data were available, there would be no need for inference), the test estimates the variance from the sample variance.

Where Δ is the paired score differences of two system-topic score lists (as in Tables 2.5 and 2.6) a t-value for the list can be computed using:

$$t = \frac{\overline{\Delta}}{sd(\Delta)/\sqrt{|\Delta|}}, \tag{2.23}$$

where $\overline{\Delta}$ is the sample mean and $sd(\Delta)$ is the sample standard deviation of Δ. A Student distribution is parameterized by the degrees of freedom parameter ν, set as $\nu = |\Delta| - 1$. With the computed t-value and ν, the cumulative density function of the Student t-distribution produces a p-value used to compute statistical significance, generally computed using a software package. As the result of that cumulative density function yields the one-sided p-value, either the p-value should be doubled to be brought into terms of a two-tailed test, or the required α threshold needed to reject the null hypothesis can be halved, exploiting the symmetry of the t-distribution. An example which evaluates the t-test is presented later in the chapter in Table 2.9 on page 54, as it is contrasted with an inferential risk overlay in Section 2.5.

Smucker et al. [179] verified the applicability of the t-test for use on IR effectiveness scores, where they measured the agreement of the p-values with the randomization and Bootstrap tests (both explained in this section), as well as the Wilcoxon and sign tests mentioned above. Smucker et al. also found that fifty topics tend to be a large enough sample size to activate the central limit theorem, with Smucker et al. [180] further noting that smaller sample sizes might still be acceptable. The *central limit theorem* roughly states that for a set of independent random variables of any shape, as the sample size $n \to \infty$, the difference distribution between the sample statistic and population statistic is described by a normal distribution. Whether the central limit theorem is believed to be in effect can play an important role in determining whether parametric inferences are valid. The activation of the theorem is also related to sample size.

Several simulation studies on IR scores among these standard testing approaches show that the t-test balances type I and type II error rates [179, 196, 197] when using the mean as the summary statistic for the test. The t-test has the added benefit that because it can be computed analytically, it can provide a p-value in sub-second time, compared to the Bootstrap

Algorithm 2.1: The Bootstrap method to compute p-values for a two-tailed null hypothesis statistical test on IR scores. As it is two-tailed, the significance required to pass the test is $\alpha/2$.

```
   Input: An array of IR effectiveness scores for an experimental system $s_1$ and
          baseline $s_2$, for $B$ bootstrap replicates, where $T$ is the statistic on the paired
          score differences.
   Output: The $p$-value for accepting the null hypothesis that $s_1$ is not significantly
           different from $s_2$, on the $T$ statistic.
 1 $\Delta \leftarrow s_1 - s_2$                                   // Array difference
 2 $t \leftarrow T(\Delta)$                                  // Statistic on original sample
                                              // Generate $B$ Bootstrap replicates for $T$
 3 upper $\leftarrow 0$
 4 lower $\leftarrow 0$
 5 for $b \in 1 \ldots B$ do
 6     $c \leftarrow$ random_sample($\Delta$, replacement=true)
 7     $t^* \leftarrow T(c)$                               // Statistic on bootstrap sample
 8     if $t^* \geq t$ then
 9         upper $\leftarrow$ upper $+ 1$
10     end
11     if $t^* \leq t$ then
12         lower $\leftarrow$ lower $+ 1$
13     end
14 end
15 return min{upper, lower}$/B$
```

and randomization tests explained below. Carterette [40] notes that it is not well-known in the IR community that the t-test is akin to performing inference on a linear regression on paired score differences between two systems.

Bootstrap Test. The Bootstrap is another null hypothesis significance test, computed by generating tens of thousands of re-samplings with replacement of the original sample to simulate a *sampling distribution.* The sampling distribution for a given statistic (for example, the median) is the distribution of sample statistics for new samples (of the same size as the original) drawn from the same population. Hence, the Bootstrap can be used to compute the confidence of point estimates for population parameters.

Algorithm 2.1 provides the algorithm for producing p-values for two-tailed hypothesis testing using the Bootstrap test. When two arrays of paired IR effectiveness scores are supplied, the paired difference is calculated using an array difference operator on line 1. Then, the provided statistic T is calculated for Δ on line 2, which is stored in the variable t that will be compared against Bootstrap resamples of Δ. The procedure for computing B bootstrap resamples is then defined on line 5, where line 6 resamples from the Δ array randomly with replacement to create a resample of the same length as Δ. With this Bootstrap resampling of Δ stored in c, the T statistic is computed on c on line 7. Lines 8 and 13 handle incrementing the count of times that the bootstrap sample statistic t^* exceeds or is less than the original sample t value. After computing these counts, the minimum of either one is returned on line 15, divided by B to yield the proportion yielding the p-value. Note that because both directions are tested, the required p-value for significance is $\alpha/2$.

Sakai [156] first proposed using the Bootstrap test to evaluate the statistical significance between two lists of IR effectiveness scores. Urbano et al. [196] evaluated the Bootstrap test's false positive rate and power, finding that it is more powerful than the t-test but exposes the researcher to a higher risk of false positives. Efron [69] extends the Bootstrap to handle skewed distributions, named the bias-corrected and accelerated Bootstrap (explored in detail in Section 3.3).

Randomization Test. The randomization test also uses re-sampling and assumes that the null hypothesis is true and centered at zero (implying no difference between the means of each sample). The key idea is to exchange the labels of the two samples and then compute the difference in means many times over, creating a sampling distribution of the difference in means. Each difference in means is tallied with respect to whether the first list is better than the second or vice versa. Finally, for each count, the p-value is computed by dividing each count by the total number of iterations. The randomization test does not assume the shape of the tails of the distribution, making it nonparametric. If all permutations are evaluated, it is an exact permutation test; but intractable to compute for long lists.

2.3.2 Power Analysis

Power analysis is a statistical methodology that helps practitioners determine the sample size needed to detect a significant difference. The power of the statistical tests can be computed analytically. Two analytical approaches are discussed below, then simulation-based procedures explored in the IR community are presented.

Webber Analytical Approach. Webber et al. [209] consider the role of statistical power analysis in IR, establishing a methodology to determine whether there is at least an 80% chance of a Student t-test yielding an outcome. They found that more than 50 topics in a statistical test with AP as the effectiveness measure in the Robust04 test collection are needed to reliably differentiate second-quartile runs against first-quartile systems. Instead, they recommend using approximately 150 topics to have adequate power. Fortunately, the Robust04 collection has 249 topics and has the largest number of deeply-judged topics available to researchers. Webber et al. [209] use the *cumulative distribution function* (CDF) Φ of the normal distribution to compute power β (not to be confused with the β distribution) for n samples:

$$\beta \approx \Phi\left(\sqrt{n} \cdot \frac{\delta}{\sigma} - z_{1-(\alpha/2)}\right) , \tag{2.24}$$

where δ is the detectable score difference sought, σ is the standard deviation of the per-topic score differences, and $1 - \alpha/2$ refers to the two-tailed quantile of the normal CDF, with $z_{1-(0.05/2)} = 1.96$. To get the number of topics required for a given (β, α, δ) combination, algebraic manipulation and the inverse CDF yields:

$$n \approx \left\lceil \left(\frac{\sigma}{\delta} \left[\Phi^{-1}(\beta) + z_{1-(\alpha/2)} \right] \right)^2 \right\rceil . \tag{2.25}$$

Webber et al. [209] also note that knowledge of the standard deviation σ of the differences in per-topic evaluation scores is a key issue. As the variance fluctuates within and between corpora, estimates of σ commonly result in pessimistically high n predictions.

Sakai Analytical Approach. Sakai [160] reexamined the experiments of Webber et al. [209] from another angle, using a method proposed by Nagata [140] to determine the sample size *a priori*. Nagata [140] employs a one-way ANOVA and assumes the system-topic evaluation scores obey a normal distribution with a joint population system mean and variance. Sakai [160] observed that the shared system variance assumption does not hold, as did Carterette [41]. Nevertheless, the one-way ANOVA is still used by IR practitioners, especially in the context of multiple testing with Tukey's honestly significant difference test (discussed soon in Section 2.3.3). The Nagata [140] method determines that n topics are required to achieve a particular significance level, power, and effect, where:

$$n \approx \left(\frac{z_{\alpha/2} - z_{1-\beta}}{min\delta_t} \right)^2 + \frac{z^2_{\alpha/2}}{2} \, , \tag{2.26}$$

in which $min\delta_t$ is the effect size, using the same notation as Equations 2.24 and 2.25. Cohen [52] proposes that the $min\delta_t$ for small, medium, and large detectable effect are 0.2, 0.5, and 0.8 respectively. The $min\delta_t$ values translate into effectiveness scores, such as δ in the Webber formulae, via:

$$min\delta_t = \frac{\delta}{\sqrt{\sigma^2}} . \tag{2.27}$$

Finally, using Sakai [160, Equation 36], the standard deviation of the per-topic score differences over the S submitted runs on T topics for an IR collection can be calculated as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{|S|} \sum_{j=1}^{|T|} (S_{ij} - \overline{S_i})^2}{|S|(|T| - 1)}} \, , \tag{2.28}$$

where S_{ij} is a system-topic score for an evaluation measure, and $\overline{S_i}$ is the sample mean evaluation score for system i.

Simulation Studies. Simulation has been used to evaluate the power of statistical tests in IR effectiveness experiments. In particular, understanding which test is the most powerful and stable is a problem IR researchers have studied for decades, with collection splitting [58, 169, 196, 202, 224] methods among the earliest approaches employed.

Smucker et al. [179] compared a range of tests with the permutation test and measured strong agreement with the t-test. While many researchers cite this study as solid support for using the t-test, Smucker et al. urged practitioners to interpret these findings with caution, as the t-test compares differences in means. In contrast, the Wilcoxon test compares differences in medians.

Parapar et al. [145] oppose the consensus that the t-test should be the preferred choice for statistical testing on IR effectiveness scores. They maintain that consistency between collection splits is not a surrogate for knowing the null hypothesis and claim that experiments that compare one test's agreement to any other test may be biased. Parapar et al. used the Manmatha et al. [126] score distribution-based approach to simulate the output of *new runs* with a fixed set of topics, asserting that generated run represents different systems, thereby fixing the null hypothesis. Parapar et al. evaluated the power of several statistical tests, finding that the Wilcoxon and sign tests had more statistical power than the permutation test.

Urbano et al. [197] rebutted the methodological soundness of treating new runs for the same topics as being analogous to controlling the null hypothesis, as the topic effect dominates IR scores, and instead proposed the simulation of *new topics* over the same set of systems. Urbano et al. used a Urbano and Nagler [195] stochastic simulation to dynamically select the mathematical distributions that best fit the score data to minimize over-fitting, and infer a copula for that marginal distribution to form new topic scores. Their results disagree with the observations of Parapar et al. [145], with the Wilcoxon and sign tests found to be the least powerful, agreeing with older work on statistical power for IR.

In an alternative approach to simulating scores to explore test outcomes, Ferro and Sanderson [73] explore the statistical properties of an ANOVA test for IR effectiveness experiments, including interaction effects of systems, topics, and shards, by randomly generating many topic sets and comparing the statistical outcomes of various tests.

2.3.3 Multiple Systems

> "...it is very difficult to conclude that *anything* is significant once we have modeled many of the sources of randomness in experimental design and analysis (on IR scores)."
>
> — Carterette [41, p. 1]

The statistical testing approach commonly used in IR ignores information about system and topic effects in favor of (nothing but) local comparisons of pairs in systems. Carterette [41] notes that two systems are seldom considered in isolation in most IR work, and oftentimes a *challenger* system is compared against multiple *champions* (baselines). As testing multiple hypotheses increases the *family-wise error*, which increases the *false-discovery rate* where the null hypothesis has been rejected wrongfully, the testing methodology should correct for this. However, Carterette found that significance between system pairs arises relatively rarely after corrective measures are applied. Carterette [41] contends that significance testing practices, as commonly applied by IR practitioners without *multiple comparison correction* (MCC), may be at least partially responsible for the stagnation of IR measurements noted by Armstrong

Algorithm 2.2: The Holm-Bonferroni multiple comparison correction method.

```
Input: An array $P$ of $p$-value outcomes from a statistical test, and the the desired $\alpha$
       significance level.
Output: An array of $(p, r)$ tuples, where $p \in P$ and $r$ is the Boolean result
        indicating whether the null hypothesis can be rejected.
1 $P \leftarrow sort(P)$                                  // Sort the $P$ array.
2 $result \leftarrow array(|P|)$                          // Allocate 1D result array
3 for $i \in 0 \ldots |P| - 1$ do
4     $reject \leftarrow P[i] < \alpha/(|P| - i + 1)$
5     $result[i] \leftarrow (P[i], reject)$               // Set result, paired with original $p$-value.
6 end
7 return $result$                                         // Return $P$ with result
```

et al. [15]. While many approaches exist to correct the family-wise error of statistical tests, the most common approaches used in IR are the Bonferroni correction and Tukey's HSD [41, 158, 162].

Bonferroni Correction. The Bonferroni correction [30] is the most straightforward multiple comparison correction technique and makes no assumptions about the distribution of the data. It is applied by dividing the significance level by the number of comparisons. For example, if the significance level is 5% ($\alpha = 0.05$) and there are m comparisons, the corrected significance level when 20 hypotheses are considered is $\alpha' = 0.05/20 = 0.0025$. Any pairwise tests exceeding this new stricter expectation are then reported as "being significant at α'". Although Bonferroni correction has many attractive properties, Sakai [158] contends that more powerful corrective techniques should be used for IR evaluation and advocates using Tukey's HSD instead.

An extension to the Bonferroni correction is the Holm-Bonferroni correction [103]. This sequential procedure tests the null hypothesis in increasing order of the individual paired p-values making up the set of m comparisons. Boytsov et al. [32] suggest using the Holm-Bonferroni method in an IR context instead of the standard Bonferroni above, claiming that it is more appropriate when m is large.

The Holm-Bonferroni method is defined in Algorithm 2.2. A set of p-values is supplied to the function in the array P, along with the significance level required α. Firstly, line 1 arranges the P array in ascending order. The resulting array is allocated on line 2. It is sequentially added to in pairs of the p-value and whether to reject the null hypothesis against this p-value. That rejection is predicated on the outcome of line 4 of the algorithm:

$$p < \frac{\alpha}{m - i + 1} \tag{2.29}$$

where p is a p-value in P, i is the rank of $p \in P$. Finally, these acceptances and rejections of the null hypothesis are returned on line 7.

Tukey's Honestly Significant Difference. Tukey's *Honestly Significant Difference* (HSD) [189] is a more powerful multiple comparison correction technique than the Bonferroni [158] one. The test is applied after an ANOVA (explained next), assuming by proxy that the data is normally distributed [92]. The key assumption in Tukey's HSD is that if the largest mean difference in a family of tests is not statistically significant, for example, between the best and worst systems, then no other mean differences will be either [41]. Hence, a studentized *range* distribution of the system mean differences are used.

When evaluating paired values with equal group sizes, Tukey's HSD is applied to the studentized range distribution [162, Equation 4.15] using the equation:

$$\frac{\overline{S_{max}} - \overline{S_{min}}}{\sqrt{(SS/\nu)/|T|}}\,, \tag{2.30}$$

where S_{max} and S_{min} are the most and least effective systems in S. The denominator brings the above *range* into studentized units [41], by calculating the sum of the squares *SS* between each of the groups of system and topic effectiveness values for the T topics:

$$SS = \sum_{i=1}^{|S|}\sum_{j=1}^{|T|}(S_{ij} - \overline{S_i} - \overline{S_j} - \overline{S})^2\,, \tag{2.31}$$

which are divided by the degrees of freedom ν:

$$\nu = (|S| - 1)(|T| - 1)\,. \tag{2.32}$$

The evaluated Equation (2.30) value is then compared to the critical value in the studentized range distribution to determine whether the null hypothesis is rejected. The interested reader is referred to Sakai [162, Section 4.4.3] for an in-depth discussion of the statistical properties of Tukey's HSD in an IR context.

ANalysis Of VAriance. *Analysis of Variance* (ANOVA) is a technique used to compare the means between two or more groups of data. It tests the null hypothesis where all groups are equal, and is commonly used in IR to compare the performance of multiple systems via effectiveness score comparisons. Similar to the Wilcoxon W table presented in Table 2.5, ANOVA is computed using an F-distribution to yield an *ANOVA table*, used to determine whether the F-value exceeds a critical F-value needed to reject the null hypothesis. In modern practice, the F and W distributions are evaluated using a statistical software package rather than by-hand lookup tables. Glass et al. [92] note that the ANOVA requires that the data is: normally distributed, the variances of the groups are equal, and the groups are independent. When only two groups are considered, the outcome of an ANOVA F-test is the same as a Student t-test [40]. Practitioners in the IR community typically use a two-way ANOVA (modeling the system and topic effects separately), which fits the linear equation to a set of observed system-topic metric scores from an experiment involving multiple topics and systems:

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_{ij} \, , \tag{2.33}$$

where i is a system associated with the system effect α_i and j is a topic connected to the topic β_j effect. Included in the linear equation is the intercept b as an error term, and the multiplied term $(\alpha\beta)_{ij}$ is an interaction term describing the joint relationship of the system and topic effect. As ANOVA deals with normally distributed response variables, the above two-way ANOVA belongs to the family of general linear mixed models (gLMMs), not to be confused with *generalized* linear mixed models (GLMMs) which are more flexible.[6]

Recent works have explored more factors in the context of IR effectiveness modeling. Sub-collections (or shards) have been studied to understand whether segmenting documents in a corpus can help to reduce statistical error [72, 171]. Zampieri et al. [219] performed a statistical analysis to determine topic difficulty, finding that the target corpus has the largest interaction effect between the set of retrieval systems, the corpus used, and system components in the retrieval model on topic difficulty. Culpepper et al. [61] recently investigated modeling topic difficulty using an ANOVA, finding that the query formulation effect is similar in importance to the topic-corpus interaction effect when explaining topic difficulty variance.

2.3.4 Bayesian Inference

> "It is not clear to me for how long the IR community will continue to embrace classical significance testing; IR researchers should be aware that Bayesian approaches to hypothesis testing are quickly gaining popularity among statisticians and among research disciplines that have heavily used statistics; we are beginning to see Bayesian approaches in the field of IR evaluation as well."
>
> — Sakai [162, p. 148]

Section 2.3.3 highlighted the challenges IR researchers face when trying to draw conclusions from their experiments involving multiple system comparisons. This section presents an overview of Bayesian inference, its current applications in IR, and its potential for future research for multiple system comparisons.

Bayesian Inference Overview. The distinguishing feature of Bayesian inference is that additional information regarding how the data is distributed is incorporated into the inference process, rather than solely deriving inferences on the frequency of an event occurring based on one or more samples. These beliefs are represented as *prior* distributions. Typical applications of Bayesian inference use weakly-informative priors, where the prior distribution has a minor influence on the posterior distribution, to ensure it conforms to an expected belief. For example, a weak prior might honor a belief that the data belongs to a normal distri-

[6] A detailed discussion of general linear models vs. generalized linear models is available at `https://stats.stackexchange.com/q/7261`. (Accessed on 13th October 2022).

	Frequentist Statistics	Bayesian Statistics
Hypothesis Test	p-value	Bayes' factor
Estimation with Uncertainty	Maximum likelihood estimate with confidence interval	Posterior distribution with credible interval

Table 2.7: An adapted table from Kruschke and Liddell [117, Figure 1], explaining the conceptual differences in Bayesian data analysis with statistical inference.

bution with a known range of location and scale parameters, a common assumption in many of the techniques discussed for both paired (Section 2.3.1) and multiple comparisons (Section 2.3.3). A more informative prior in the context of a particular IR effectiveness metric could be a belief that the distribution of scores is constrained to lie in the interval $[0, 1]$. In a weakly-informative setting, the observed data is the primary source of information, and the parameters in the prior distribution (such as the location μ in a Gaussian distribution) are updated to fit the data during the inference process better.

Table 2.7 (adapted from Kruschke and Liddell [117, Figure 1]) explains the difference in data analysis techniques applied across the inferential paradigms, where the bottom-right quadrant using the posterior distribution in a manner similar to how it is used by Gelman and Shalizi [86] is the space explored in this thesis. One method of measuring the strength of the null hypothesis in a graded way is to use Bayes' factors. Bayes' factors measure how well one model fits the data over another.[7] A point-wise estimate of the null hypothesis distribution (in this case $H_0 = 0$, that is, that there is no difference in the means of the two sets of values) is used to compare the density ratio of the Bayesian prior and posterior distributions. In a model selection problem for text classification, Zhang et al. [220] used Bayesian inference to make inferential claims using graded Bayes' factors and *precision*-based approaches with credible intervals. (Not to be conflated with the IR definition of precision.) An alternative to using Bayes' factors is measuring the precision of an estimated model parameter [115]. That estimate is usually supplemented with a 95% *highest density interval* (HDI) and whether zero is included in that region. Significance testing in a Bayesian sense with multiple groups involves estimating an effect and providing uncertainty intervals about that estimate. Bayesian inferences in this thesis are applied with respect to the precision of an estimate.

Although there are cases where Bayesian inferences can be computed without simulation, these are limited to situations where the posterior distribution can be expressed using the same distributional family as the prior, known as a *conjugate prior*. *Markov chain Monte Carlo* (MCMC) is a technique that enables the simulation of posterior distributions with greater flexibility. Figure 2.5 on the next page overviews the steps required to perform Bayesian inference using MCMC methods. At a high level, a researcher will provide a set of inputs to an MCMC sampler. If the sampler can complete the sampling process, the output is then validated using a set of diagnostics. The researcher can make inferences about the posterior distribution if the

[7] The `bayestestR` package provides a useful introduction at `https://easystats.github.io/bayestestR/articles/bayes_factors.html`. (Accessed on 13th October 2022).

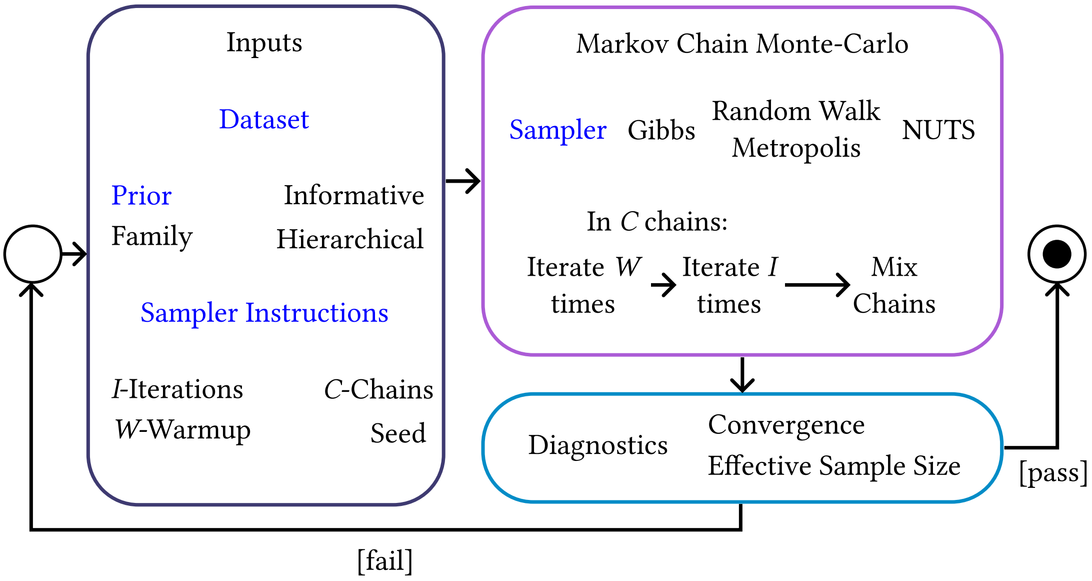


Figure 2.5: The process involved in Markov chain Monte Carlo (MCMC) sampling for Bayesian inference. Given a set of inputs and an MCMC sampler, if the sampler can complete the sampling process, the output is then validated using a set of diagnostics. If the diagnostics indicate sampling errors, the researcher must start again with a different set of choices. High-level researcher decisions are labelled in blue text, where sub-decisions are placed adjacent to the blue labels.

diagnostics pass. Otherwise, the researcher must start again and evaluate why the sampler failed. The observed data is the primary information source for the inference process. The researcher must make modeling decisions depending on the number of observations. Too few observations may not provide enough information for the Markov chains to converge due to lack of evidence, and too many observations may result in long completion times for more complicated models.

Algorithm 2.3 on the following page presents pseudo-code for the MCMC algorithm, taking as inputs: the observed data, a set of parameters describing the data, a log-density function for calculating the probability of seeing the observed data with the set of parameters, and MCMC settings as inputs. The MCMC settings describe the count of iterations to run, the number of chains to execute, and the number of warm-up iterations to discard. (The warm-up iterations allow the Markov chains to converge to the posterior distribution and are not used for inferential purposes.) Each chain can be run asynchronously, where each chain samples the posterior with respect to the parameter set generated by the previous iteration. The accepted parameter set for a particular chain and iteration combination is mixed into a single one-dimensional array. Analyzing the results of particular chains is typically only done when diagnosing convergence issues. In practice, the more chains there are, the less chance the sampler will be constrained in a local minimum and producing misleading inferences. Lambert [118, p. 314] recommends 4–8 chains for simple models, and "few tens of chains"

Algorithm 2.3: The general steps for performing Markov chain Monte Carlo (MCMC) to compute the posterior distribution used for inference. Chain computations are often parallelized, joining each thread before returning the result.

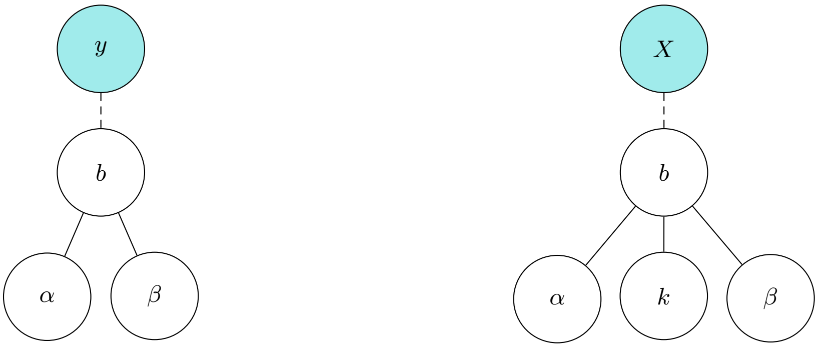
```
Input: The array $Y$ of data observations, the set of parameters to model $\theta^p$, the
       number of MCMC chains $C$, iterations $I$, warmup iterations $W$, and the
       log-density function lp_fn used to calculate the probability of an iteration's
       parameter set describing the data. The sampler_propose and sampler_decide
       functions are specific to the sampler used.
Output: The posterior distribution $\theta$ of the parameters $\theta^p$ describing $Y$.
$\theta \leftarrow array(C \times (I - W))$                          // Initialize 1D array
for $c \in 0 \ldots C - 1$ do
    $prev \leftarrow initialize\_params(\theta^p)$                   // Random values
    for $i \in 0 \ldots I - 1$ do
        $prop \leftarrow sampler\_propose(lp\_fn, Y, prev)$          // Propose new parameter set
        $prev \leftarrow sampler\_decide(prev, prop)$                // Accept or reject
        if $i > W - 1$ then
            $\theta[((i - W) \times C) + c] \leftarrow prev$         // Save non-warmup proposals
        end
    end
end
return $\theta$                                                  // Return mixed posterior distribution
```

for complex models to mitigate against the MCMC sampler settling in local minima. The *initialize_params* function assigns a random number to each parameter in the model, in Stan[8] and other Bayesian inference implementations [147].

There are many MCMC samplers available, and the appropriate choice of the sampler can depend on the complexity of the model, the number of observations, and the researcher's tolerance for the efficiency vs. accuracy trade-off during sampling. As the MCMC sampler runs, it will produce a set of *dependent* samples from the posterior distribution. The samples are then validated using a set of diagnostics to ensure the sampling process is successful, often checking for chain convergence while verifying that the number of samples produced is sufficient for the researcher's goals. For example, to support a 95% HDI, Kruschke [116] recommends a minimum effective sample size of 10,000. The $\hat{R}$ heuristic [85] measures how well the Markov chains computed have mixed where values closest to 1.0 are most favorable.

A popular modern MCMC sampler is the *No-U-Turn Sampler* (NUTS) [101]. The No-U-Turn-Sampler works by inverting the posterior distribution (transforming its shape from a hill to a valley) and then using Hamiltonian dynamics to sample whether to move up or down the valley. Each step in the sampling process is a random walk. The NUTS algorithm uses the log-density of the model (given the observed data and set of candidate parameters) to determine whether the sampler should stay in the current position or move to a new position. Lambert [118] uses the analogy that the NUTS sampler is akin to a snow sledge (or toboggan) gliding over a frictionless valley to explore the area, with gravity dictating that the sledge

[8] Via `rstan` documentation: `https://search.r-project.org/CRAN/refmans/rstan/html/stan.html`. (Accessed on 11th November 2022).



(a) Typical IR inferential model.

(b) Modeling relevance at document-rank level.

Figure 2.6: The Bayesian models Carterette [42] explored when modeling IR scores, denoted y. The system effect is represented by α, and the topic effect is denoted as β, where any error not ascribed to these effects is propagated into an error term. An alternative to modeling IR effectiveness scores for measuring system dominance is to model the system-topic-rank relevance judgment values as a random variable X and use a GLMM for inference.

will move down the valley and spend more time in the lower regions. The lowest point in the valley is the most likely position for the parameters to explain the observed data given the model. For an algorithmic description of the NUTS sampler, see Hoffman and Gelman [101].

A full explanation of Bayesian data analysis is beyond the scope of this thesis. Many excellent resources are available for those interested in learning more about Bayesian data analysis, including the following:

- Lambert [118] provides a comprehensive introduction to Bayesian inference approachable for a general audience, and is up-to-date with recent developments in the field.
- Kruschke [116] provides a complementary resource with different examples to help the reader understand the concepts.
- Gelman et al. [89] has a more technical mathematical focus, and can be used as a reference for a range of advanced concepts after reading the two introductory texts.

Bayesian Inference in IR. Carterette [40] pioneered early work on Bayesian inference, demonstrating that the methods are able to produce deductions similar to the linear models such as the Student t-test and ANOVA. Additionally, modeling relevance directly to measure relative system effectiveness rather than computing evaluation metrics on these values is also explored. The key goal of the work was to "begin to form alternative models for evaluation that more precisely capture aspects of IR effectiveness that are not captured by the t-test" [40, p. 4]. This work was later extended by Carterette [42] to compare the p-values generated by the Student t-test and various Bayesian models. In both papers, the posterior was simulated using MCMC with *Just Another Gibbs Sampler* (JAGS) [146] sampling.

Figure 2.6 shows the Bayesian models Carterette [42] explored when modeling IR scores. On the left is the typical modeling approach usually considered in an IR context, where the system effect is represented by α, the topic effect is denoted as β, and where any error not

Algorithm 2.4: The posterior log-density accumulator based on Toyoda [187] explored by Sakai [161] for hypothesis testing on paired IR effectiveness scores.

```
Input: The two-dimensional array $Z$ of IR effectiveness score *pairs* for two systems,
       with current chain (or, initial) proposals for $\langle \mu, \sigma, \rho \rangle$ used to explore the
       posterior distribution.
Output: The accumulated log-posterior density for the given inputs, for the MCMC
        sampler to probabilistically determine whether remaining in the current
        position or moving to this location is optimal.
$lp \leftarrow 0$          // The absence of prior density accumulation indicates a uniform prior
$v\_cov \leftarrow array(2, 2)$                      // Initialize 2D variance-covariance matrix
$v\_cov[1][1] \leftarrow \sigma[1]^2$          // Leading diagonal represents variance of each system
$v\_cov[2][2] \leftarrow \sigma[2]^2$
$cov \leftarrow v\_cov[1][1] \times v\_cov[2][2] \times \rho$        // Model covariance using the proposed $\rho$
$v\_cov[1][2] \leftarrow cov$                        // Trailing diagonal represents covariance
$v\_cov[2][1] \leftarrow cov$
for $z \in Z$ do
    $lp \leftarrow lp + multi\_normal\_lpdf(z, \mu, v\_cov)$        // Accumulate likelihood density
end
return $lp$
```

ascribed to these effects is propagated into an error term. An alternative is to model the system-topic-rank relevance judgment values as a random variable X and use a GLMM to explore the system effect on these values.

In another study, Sakai [161] described how using a Bayesian t-test can be more informative than the classic Student t-test and how the measures of interest to IR researchers can be extracted from simulated *posterior predictive distributions* (PPDs). Sakai [161] used the Stan probabilistic programming language [39] to simulate the posterior distribution, using NUTS [101] to perform MCMC. Glass' delta was suggested to provide useful context around the effect size when reporting inferential results, as p-values can be misleading due to the relationship with sample size [159]. Sakai [161] explains that Bayesian inference is an accepted form of practice in the statistical community, but it has not to date been widely used in IR.

Algorithm 2.4 presents the Bayesian paired modeling approach in detail, based on the presentation given by Toyoda [187].[9] In this approach, the NUTS sampler is provided with the log-density function of the parameters and model specification. The sampler is used to generate the posterior distribution by sampling from it. A multivariate normal distribution is used to model the effectiveness of the systems, and ρ is used to model the correlation between the two distributions. Sakai [161] notes that the posterior distribution is computed using a uniform prior (without an explicit prior specification), which is reflected in Algorithm 2.4. (Examples of posteriors with non-uniform priors are described in Chapters 4 and 5.) Al-

[9] The Stan code for the paired modeling approach presented in Sakai [161] was taken from Professor Tetsuya Sakai's website at `http://sakailab.com/download/`. (Accessed on 13th October 2022).

though the sampler could have been supplied with the paired score differences as the data for inferential purposes with only μ and σ simulated, the ability to model the correlation between the two systems is one example showcasing the flexibility of Bayesian modeling.

Generalized linear mixed models in combination with Bayesian inference have been applied in various IR statistical analyses. Alanazi et al. [11] used a GLMM for binomially distributed values with a logit function, studying whether ad quality and position influence user satisfaction, search time, and behavior in a lab-based eye-tracking user study. Checco et al. [46] used GLMMs to model the agreement of relevance assessors in a crowdsourcing experiment using a Beta distribution. Urbano and Nagler [195] explored modeling various IR effectiveness metrics with tighter constraints about their response distributions. For example, P@10 has a discrete distribution with 11 possible values in the range $\{0.0, 0.1, \ldots, 1.0\}$, however, most IR statistical analyses treat effectiveness scores as belonging to a continuous distribution. Urbano and Nagler found in their study that it was rare for one model to be uniquely the one that provides the best fit across a set of observed IR scores. Saitoaki et al. [155] use Bayesian inference and GLMMs to describe searcher behavior when weasel sentences are highlighted on a webpage using the `brms` `R` package instead of a traditional ANOVA analysis. Time statistics were modeled using Weibull distributions, while page views were abstracted as belonging to a Poisson distribution for count data.

2.4 Statistical Distributions

This section describes the main statistical distributions used in this thesis. These distributions are used to model IR effectiveness data and to generate inferences. Many other distributions should thus be considered (see Section 5.1 for a holistic empirical comparison of many distributions), but the ones summarized here are the most important.

2.4.1 Gaussian

The Gaussian distribution, also known as the normal distribution, is a continuous probability distribution. It is the most common statistical distribution and has many advantageous properties, representing data in the range $(-\infty, \infty)$. The Gaussian distribution has two parameters, the mean μ and the standard deviation σ. The mean is a measure of the central tendency of the distribution, and the standard deviation describes the spread of the data symmetric around the mean. The *standard* Gaussian distribution has a mean of 0 and a standard deviation of 1. Gaussian distributions have one mode, where the mean, median, and mode are equal. The Gaussian probability density function is defined as [76, p. 143]:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} . \tag{2.34}$$

	Location	Scale
Gaussian	μ	σ^2
Skew-Normal	$\xi = \mu + \sigma \frac{\lambda}{\sqrt{1+\lambda^2}} \sqrt{2/\pi}$	$\omega = \sigma^2 \left(1 - \frac{2\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2}{\pi} \right)$

Table 2.8: Differences in the location and scale parameters for the Gaussian distribution against a Skew-Normal distribution.

2.4.2 Skew-Normal

The Skew-Normal [16, 76, 143] distribution extends the Gaussian distribution to additionally model skewness. In some circumstances, data is not symmetric around the mean. The Skew-Normal distribution can be used to describe that asymmetry while more accurately reflecting the central tendency of the data.

The Skew-Normal has the definition:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \times \left[1 + erf\left(\lambda \frac{\frac{x-\mu}{\sigma}}{\sqrt{2}} \right) \right] , \tag{2.35}$$

where *erf* is the Gauss error function [213], defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt \, . \tag{2.36}$$

While the location and scale of the Gaussian distribution is defined in terms of the mean and variance, the Skew-Normal distribution describes location ξ, scale ω, and *shape* λ in terms of the Gaussian mean μ and variance σ^2. However, because the location of a Skew-Normal distribution considers the skewness parameter, the "mean" of the Skew-Normal distribution is not particularly useful for comparative purposes, but rather the location parameter ξ is more useful. A similar argument can be made for the scale parameter ω against the variance σ^2. (Unless, as noted by Azzalini [16], if the skewness is modeled to be 0, then the Gaussian and Skew-Normal means are equivalent.)

Table 2.8 compares the location and scale parameters of the Gaussian against the Skew-Normal distributions. Note that the location parameter ξ of the Skew-Normal distribution is not only bound to the coordinates of the Gaussian mean μ, but it is also adjusted in terms of the joint relationship between the variance and the skewness. Additionally, the scale ω of the Skew-Normal distribution takes the original variance σ^2 of the Gaussian distribution and adjusts it in terms of the skewness λ. Therefore, the location and scale parameters of the Skew-Normal distribution are similar to the location and scale of the Gaussian distribution, but they are adjusted to account for the skewness of the underlying data.

2.4.3 Beta

The Beta distribution represents values in the interval $(0, 1)$, which is parameterized by two shape parameters, α and β. Due to its bounded nature, the shape parameters allow for more flexible modeling of values between zero and one than is possible using the Gaussian distribution.[10] The Beta distribution is defined as [76, p. 55]:

$$f(x) = B(\alpha, \beta)^{-1} \times x^{\alpha-1}(1-x)^{\beta-1}\,, \tag{2.37}$$

where $B(\alpha, \beta)$ is the Beta function:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}\,(1-t)^{\beta-1}\;dt\,. \tag{2.38}$$

As the Beta distribution is defined with the open interval range $(0, 1)$, zero and one cannot be modeled using it. The Zero-One-Inflated Beta distribution (ZOiB) extends the Beta distribution to include these values [144]. For the ZOiB, if an observed value is a zero, the probability of observing a zero value is recorded as a Bernoulli trial. Likewise, if the observed score is one, that too is registered in the model as a Bernoulli trial. Otherwise, the value is modeled as it would have been using a standard Beta distribution.

2.4.4 Visualizing The Distributions

Figure 2.7 shows the probability density functions of the Gaussian, Skew-Normal, and Beta distributions, with varying parameters. So that the distributions are comparable to the acceptable range of the Beta, the standard Gaussian distribution displayed on the left-most graph is translated right by 0.5. The middle plot displays the Skew-Normal distribution. Observe that when the skewness parameter λ is 0, and the same transformations are applied to the location (mean) and scale (standard deviation) parameters, the distribution is equivalent to the Gaussian distribution. Finally, in the right-most plot, the Beta distribution is displayed with varying shape parameters α and β, demonstrating its flexibility in modeling ratio values.

2.5 Risk Measures

Many risks must be managed when an IR system is actively resolving the information needs of searchers globally. The most obvious risk is that the system will cease to provide value to the user and impact advertising revenue. That risk can be mitigated by managing the components Frøkjær et al. [80] define as prerequisites for a usable IR system: effectiveness, efficiency, and satisfaction.

For effectiveness, the risk is that the system fails to provide relevant results against queries; and for efficiency, the system fails to provide results fast enough. Satisfaction might be impacted by perceptions of uptime [215], fairness and transparency [44], and other factors

[10] An animated visualization the different shapes supported by the Beta distribution can be found at `https://en.wikipedia.org/wiki/File:PDF_of_the_Beta_distribution.gif`. (Accessed on 14th October 2022).

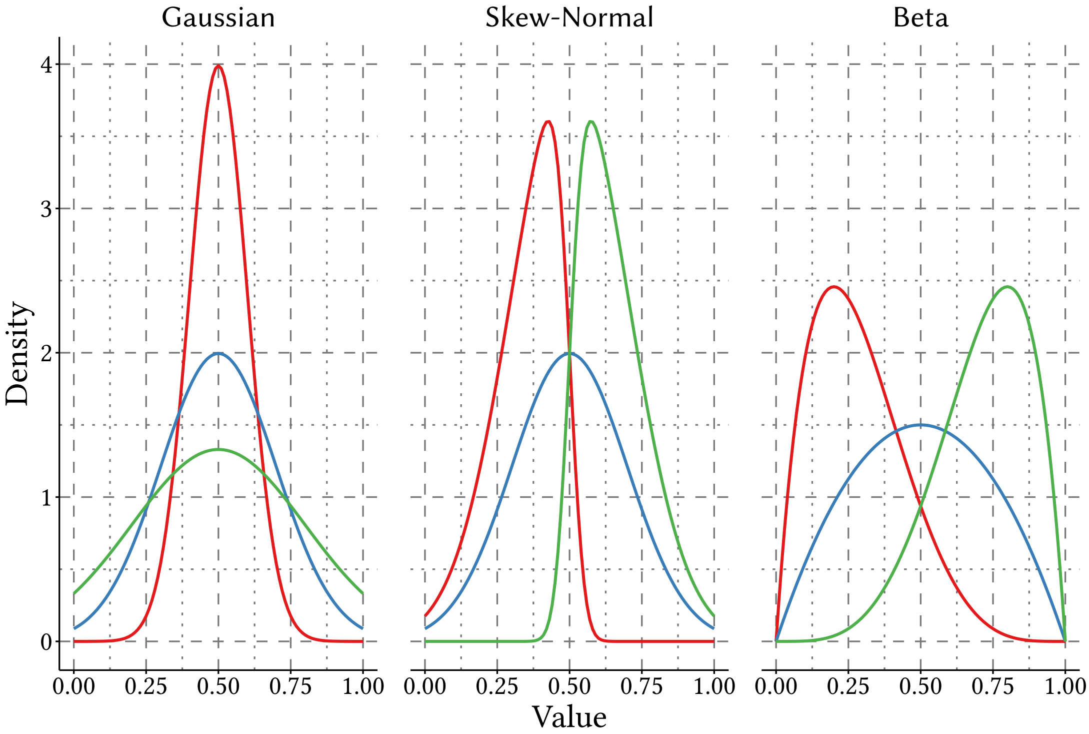


Figure 2.7: The different distributional shapes that can be modelled by Gaussian, Skew-Normal, and Beta distributions. For the Gaussian distribution, the mean is fixed at 0.5 and the standard deviation varies between 0.1 (red), 0.2 (blue), and 0.3 (green): colors follow the same order for the next set of parameters. For the Skew-Normal distribution, the location and scale are fixed at 0.5 and 0.2, respectively, and the skewness is set to either -5, 0, or 5. The Beta distribution has two shape parameters, α and β, which are set to either 2 and 5, 2 and 2, or 5 and 2 respectively.

such as color palettes, fonts, and the speed of initial page load. From a business risk perspective, Loewerre and Dominiquini [121] note that failure to innovate provides competitors with windows of opportunity to gain market share.

This thesis focuses on statistical techniques which detect IR systems with unpredictable effectiveness profiles and provide probabilistic risk assessments for a supplied risk appetite. An IR system with high mean effectiveness may also yield highly variable effectiveness across topics. For example, suppose that an IR system previously retrieved relevant results on popular topics and does not after transitioning to a new ranker. In that case, the search engine cannot resolve these information needs, where it was previously able to be relied upon to do that. Given the high-stakes nature of transitioning to a new ranking algorithm, generally, engineers will use an A/B test to segment a small percentage of users towards using the new ranker, contrasted against the main production deployment. Although A/B testing is a common due diligence strategy across the software industry, using an offline diagnostic tool before exposing even a tiny percentage of users to a new system is a prudent first step. As searchers in an A/B group are exposed to either the champion or challenger ranker blindly, their perception of the quality of the search engine may be permanently damaged when served an ineffective ranking, placing more faith in a competing product from therein. As many of the proposed IR risk measures have borrowed from economics, econometrics measuring market volatility are first reviewed, followed by risk measures in IR.

2.5.1 Risk Measures in Economics

Risk Psychology In Economics. Prospect theory is a model in behavioral economics which asserts that people behave irrationally under conditions of uncertainty. Tversky and Kahneman [194] show that people underweight likely outcomes that are the better option in favour of absolute certainty, while preferring unlikely outcomes when both options are considered a loss. On balance, the theory explains that people are inherently risk-averse and experience more pain from losses than satisfaction from gains of the same magnitude.

Prospect theory is composed of two functions. The first is the value function $v(x)$, which ascribes the estimated psychic gain or loss a person places on an outcome relative to a reference point. The second function is the weighting function $\pi(p)$, where each value function outcome is multiplied by a decision weight when the probability of the event p is known. Abdellaoui and Kemel [5] investigate the relationship between gains and losses of time, as opposed to money. Those authors show that the perception of losing time in the short-term (0–60 minutes) is different to money; however, prospect theory is still a useful model if the value function is tuned to honor these findings.

Quantitative Risk In Economics. Maiti [124] provides a recent survey into the use of risk measures in economics. They identify four major areas that have influenced the development of risk measures for economics:

- Markowitz [128] introduced the concept of modern portfolio theory.
- Sharpe [174] introduced work on a capital asset pricing model (CAPM).
- Fama and French [70] introduced the Fama-French three-factor model, later publishing the Fama and French [71] five-factor model.

Modern portfolio theory is a methodology that combines both the mean and variance in an asset price to determine whether it should be included in a portfolio, given an investor's risk appetite. Given two portfolios where the expected return is known to be identical, the portfolio with lower variance is considered the less-risky option. The capital asset pricing model extends portfolio theory to model the volatility of an asset by comparing it with a representative sample of assets in the market.

The Fama-French three-factor model is a regression derived from historical market data. The key observations are that value stocks outperform growth stocks, smaller companies outperform larger companies, and stocks with a high book value (assets minus liabilities) to market value ratio perform the best. A value stock is a security trading at a discount, where its innate value is predicted to be higher than the current market price. In contrast, a growth stock is an asset expected to outperform the market in the future. The Fama-French five-factor model extends the three-factor model by adding that stocks reporting higher earnings growth outperform those that are not. Meanwhile, companies that are reinvesting their profits into obtaining more assets tend to have lower returns than companies that are not.

2.5.2 Risk Measures in IR

Concretely, risk in IR relates to user abandonment in search engines. The first investigation into *overlays* that capture the downside risk across multiple topics instead of purely comparing means was established by Collins-Thompson [53]. Collins-Thompson proposed the *R-Loss* at k metric, which returns a count of the relevant documents lost between an experimental system and a baseline, using it to quantify the risk of various query expansion settings. They plot R-Loss against the percentage AP gain, where the gains in effectiveness are placed in context with the count of relevant documents missing from the challenger system against the champion. If the Pareto frontier of the challenger approach is geometrically above another challenging system on the said graph, it is the superior choice on the pair of measures. Collins-Thompson [53] refers to the economics concept of the *efficient frontier*, where the optimal risk-reward tradeoff is formed on a curve on the upper-left boundary of the risk vs. effectiveness plots. This approach was later applied by Collins-Thompson [54] to evaluate their constrained optimization method for tuning query expansion parameters.

In borrowing from modern portfolio theory (discussed above), Wang and Zhuhan [206] explore how mean-variance analysis can be utilized to tune parameters within a retrieval model to minimize the overall risk of the system. Modern portfolio theory is further used by Zhu et al. [223] to form a risk-aware language modelling framework that adjusts the ranker parameters to match the personalized risk appetite of searchers. Both papers use *k-call* as a risk evaluation metric, an approach proposed by Chen and Karger [47]. The k-call measure observes top-10 rankings, where the result is either 0 or 1. For example, a value of 1 is returned if k relevant documents exist in the top-10 set, a value of 0 is assigned otherwise.

Although k-call and R-Loss had both been used to quantify risk, these measures work in tandem with a conventional effectiveness metric. Wang et al. [207] transition from graphical risk analysis and introduce an analytical method originally denoted T_α for *trade-off*, later known as URisk in the 2013 and 2014 TREC web tracks [55, 56].

To calculate URisk, allow a challenger system s_1 to be compared against the champion system s_2. Where s_1 and s_2 are ordered lists of IR effectiveness values on a pair of systems on the same corpus with the same T topics, let $\Delta = s_1 - s_2$ be the set of paired score differences. With that, the URisk overlay is defined as:

$$\mathit{URisk}_\alpha = \frac{1}{|T|}\left[\sum_{j\in T^+}\Delta_j - (1+\alpha)\cdot\sum_{j\in T^-}\Delta_j\right], \tag{2.39}$$

	Topics														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
s_1	0.4	0.5	0.1	0.2	0.2	0.2	0.0	0.4	0.3	0.0	0.5	0.1	0.4	0.0	0.1
s_2	0.1	0.1	0.7	0.8	0.6	0.6	0.7	0.3	0.3	0.8	0.7	0.1	0.5	0.1	0.8
Δ	$+0.3$	$+0.4$	-0.6	-0.6	-0.4	-0.4	-0.7	$+0.1$	0.0	-0.1	-0.2	0.0	-0.1	-0.1	-0.7
Δ_5	$+0.3$	$+0.4$	-3.0	-3.0	-2.0	-2.0	-3.5	$+0.1$	0.0	-0.5	-1.0	0.0	-0.5	-0.5	-3.5

		Δ (Standard t-test)			Δ_5 (*TRisk*$_4$)		
$\overline{\Delta}$	$\overline{\Delta_5}$ (*URisk*$_4$)	*sd*	t	p-value	*sd*	t	p-value
-0.253	-1.480	0.380	-2.585	0.022	1.590	-3.605	0.003

Table 2.9: A worked example for calculating a standard Student t-test and the URisk and TRisk overlays where losses are scaled fivefold. The t-value is computed using Equation (2.23). The R implementation of the cumulative density function pt is used to convert the t-value into a p-value, where the degrees of freedom are set to $\nu = 15 - 1$, and the output of pt is doubled to yield a two-sided p-value. As *URisk*$_4$ is negative, and *TRisk*$_4$ is negative with statistical significance, s_1 risky to swap s_2 with when losses are counted five times as important as gains. Even when no risk calculation is applied, s_1 is significantly less effective than s_2.

where T^+ indicates topics that s_1 has outperformed s_2 on, and conversely, T^- are the topics where the champion system is more effective than the challenger. The α parameter enables a practitioner to weigh these losses more heavily. Note that when $\alpha = 0$ in this scenario, the URisk equation represents the arithmetic mean.

The values generated by the URisk calculation are descriptive. These values only describe the observed sample and do not provide any information about how risky the system may be until they are placed in context with other systems. Dinçer et al. [67] introduce the TRisk overlay, translating the URisk overlay into an inferential one by studentizing the URisk values. Methods by which samples of effectiveness scores can be studentized were explained in Section 2.3, where Table 2.9 provides a worked example of the traditional Student-t against TRisk, where $\alpha = 4$ (or losses are scaled fivefold). First, the differences in effectiveness for each topic pair are evaluated in Δ. Then to calculate URisk, the subzero score differences are multiplied by the risk factor. The mean difference is calculated between the systems, as well as the risk-adjusted row Δ_5 to calculate *URisk*$_4$.

Both Δ and Δ_5 can be studentized to yield an inferential statistic which can be tested using the Student t-test. The sample standard deviation is supplied to Equation (2.23) on page 35 to calculate the t-value for the standard t-test computation, and the URisk adjusted values to calculate TRisk. When 50 topics are considered using the TRisk overlay, scores less than -2 indicate that a challenger system has presented a statistically significant risk of harming the existing champion system. In contrast, t-values exceeding 2 indicate that the experimental method can be ruled out as having a significant risk impact. For all other values, interpreting significance is inconclusive. However, the magnitude provides a measure of the risk impact. A more exact approach when the topic count is different to 50, as in the example presented in Table 2.9, is to compute the cumulative density function using a statistical software package, supplying $\nu = |\Delta| - 1$ as the degrees of freedom. The standard t-test and the TRisk approach find the differences statistically significant; the t-test interpretation is that the effectiveness of either list is significantly different, whereas the TRisk guidance is that it is risky to use s_1 over s_2 with the $\alpha = 4$ risk parameter.

While the URisk and TRisk methods are valuable tools for failure analysis, they only assess the risk of pairs of systems simultaneously. A matrix of system-topic values may be preferred if the goal is to observe the changes in risk sensitivity against many systems. Dinçer et al. [65] address this by introducing the ZRisk measure, which uses the Chi-squared statistic to provide a risk value over many baseline systems. It has a similar definition to URisk:

$$\mathit{ZRisk}_\alpha(S_i) = \frac{1}{|T|}\left[\sum_{j \in T^+} z_{ij} + (1+\alpha)\sum_{j \in T^-} z_{ij}\right], \tag{2.40}$$

where i is a system in the entire set of systems S evaluated using the overlay, and T^+ are the set of topics issued to each system in S where $z_{ij} > 0$, and T^- are the set of topics where $z_{ij} < 0$. The z_i set refers to transformed set of z-values of the system i on the S_i scores, in connection with the whole matrix of system-topic scores S. For a given system-topic combination (i, j), where $j \in T$ in the S matrix, the value is standardized using the equation:

$$z_{ij} = \frac{S_{ij} - e_{ij}}{\sqrt{e_{ij}}}, \tag{2.41}$$

and e_{ij} is the expected value of S_{ij} against the systems and topics supplied to ZRisk:

$$e_{ij} = \frac{S_i \cdot S_j}{|S| \cdot |T|}, \tag{2.42}$$

where S_j is the set of effectiveness scores on the query j for any system in S.

The z-value can be used similarly to the t-value; however, it measures the goodness-of-fit of the experimental system against the baseline systems instead of the effectiveness difference between them. Dinçer et al. [65] note that ZRisk is sensitive to variance and the shape/form of the approach under test. Yet, it does not capture information about the overall effectiveness like the previous metrics. Hence, GeoRisk was introduced in the same paper, taking the geometric mean of the ZRisk scores and the original effectiveness metric to capture information on all three goals. GeoRisk has the following definition:

$$\mathit{GeoRisk}_\alpha(S_i) = \sqrt{\overline{S_i} \cdot \Phi\left(\overline{\mathit{ZRisk}_\alpha(S_i)}\right)}, \tag{2.43}$$

where Φ is the cumulative distribution function of the standard normal distribution.

		Sensitivity			Scope		
Measure	Type	Mean	Var.	Shape	Baseline(s)	Penalty	Main Property
AP (arithmetic)	Descriptive	✔	✗	✗	✗	✗	✗
AP (geometric)	Descriptive	✔	✗	✗	✗	Low score bias	✗
k-call	Descriptive	✗	✗	✗	✗	k rel. docs = T, $\neg k$ = F.	✗
R-Loss@k	Descriptive	✗	✗	✗	One	N	Mean
URisk	Descriptive	✔	✗	✗	One	r	Mean
TRisk	Inferential	✔	✔	✗	One	r	Mean
ZRisk	Inferential	✗	✔	✔	All	r	Shape
GeoRisk	Descriptive	✔	✔	✔	All	r	Mean
Chapter 3 (BCa)	Inferential	✔	✔	✔	One	r	Mean
Chapter 4 (BRisk)	Inferential (MCC)	✔	✔	✔	Many	r	Mean
Chapter 5 (PPDRisk)	Inferential (MCC)	✔	✔	✔	Many	r	Mean

Table 2.10: Research gaps in risk-sensitive evaluation measures that have been used to compare different retrieval models, expanding on Dinçer et al. [65, Table 1].

2.6 Research Gaps

Table 2.10 summarizes the research gaps in risk-sensitive evaluation measures, expanding on Dinçer et al. [65, Table 1]. The sensitivity columns describes the statistical quantities that each measure honors, and the scope columns describe key features of each measure. The top partition of the table shows the measures that have been used in prior work, and the bottom part tabulates the proposed methods in this thesis in Chapters 3 to 5. The measures are ordered by their evolution over time.

A key trend in the development of risk-sensitive evaluation overlays are the introduction of measures that facilitate the comparison of multiple systems. In addition to wanting to compare many systems, it is also desirable to be able to make inferences about risk when analyzing the outcomes of experiments. When the GeoRisk overlay is used to compare many systems, the inferential component of ZRisk is removed and GeoRisk becomes descriptive. This thesis examines methods to make risk inferences across many systems.

In addressing the gap where there is no paired inferential risk-sensitive evaluation overlay that honors the location, variance, and shape of the risk-adjusted score distribution, Chapter 3 introduces BCa, with its inferential outcomes compared against the TRisk measure. The need for honoring the shape is further validated by Bayesian modeling in Chapter 4.

Comparing many systems is a common situation in IR, where multiple comparison correction methods are used to avoid drawing false conclusions. The false-positive rate is a factor that increases on each subsequent comparison that is made. With that, risk-adjusted score distributions may differ from standard IR scores and may not be amenable to traditional corrective methods. To address that gap in the literature, Chapter 4 introduces BRisk, a Bayesian risk-sensitive evaluation approach that models the risk of many systems with a risk-transformation pre-applied and corrects for the false-discovery rate. In deriving BRisk, Chapter 4 also introduces the first Bayesian inferential evaluation scenario over multiple systems of standard IR evaluation values. Chapter 5 presents PPDRisk, which models systems with their unadjusted effectiveness scores and applies a risk transformation to the posterior predictive distribution of the model for inferential analysis. This approach is more flexible than BRisk, as it allows for many different summary statistics to be computed, and it is statistically more powerful than BRisk.

Since results in this thesis compare the detection sensitivity of different statistical tests, the look elsewhere effect is of great concern and restricts the generalizability of the results to other evaluation measures without further investigation. Webber [212, p. 140] warns in the context of the look elsewhere effect, "trying one test collection (or metric) after another until some meaningful outcome is achieved is not, to say the least, methodologically sound." Therefore, because the combined result of the look elsewhere effect against IR metrics, statistical tests, and risk-sensitive evaluation measures is unknown, the results of this thesis are restricted to the AP and RBP metrics to avoid false rejections of the null hypothesis. The AP measure collects relevance information about the entire ranked list (recall-oriented), whereas RBP is more sensitive to the head of the result list (precision-oriented). This does not dispute the soundness of broader investigations of other contexts in the literature concerning existing IR metrics and statistical tests combinations. If this thesis demonstrates the superiority of one measure over the other in terms of detection sensitivity in a risk-sensitive evaluation scenario, then an important caveat must of necessity be that the same conclusion may not hold for other measures and more investigation is required.

3

Extending Paired Inferential Risk Overlays

Determining when to swap an existing ranker with an improved alternative is a fundamental problem in IR. As an example of this question, consider Figure 3.1, which demonstrates the variance in system effectiveness by plotting the monotonically decreasing NDCG@10 scores of a BM25 ranker against a different “RCC” approach (explained by Benham et al. [24]) on 500 queries. Comparing the mean effectiveness of each ranker, BM25 and RCC have mean NDCG@10 scores of 0.170 and 0.263 respectively; passing a t-test for $p < 0.05$. Although points above the BM25 reference line confirm that the RCC ranker is more effective in overall NDCG magnitude, there are many concerning drops in effectiveness below the reference line. These reductions in ranking performance come with a *risk* that searchers who previously relied on answers to these topics may abandon their session at best, and at worst, have a

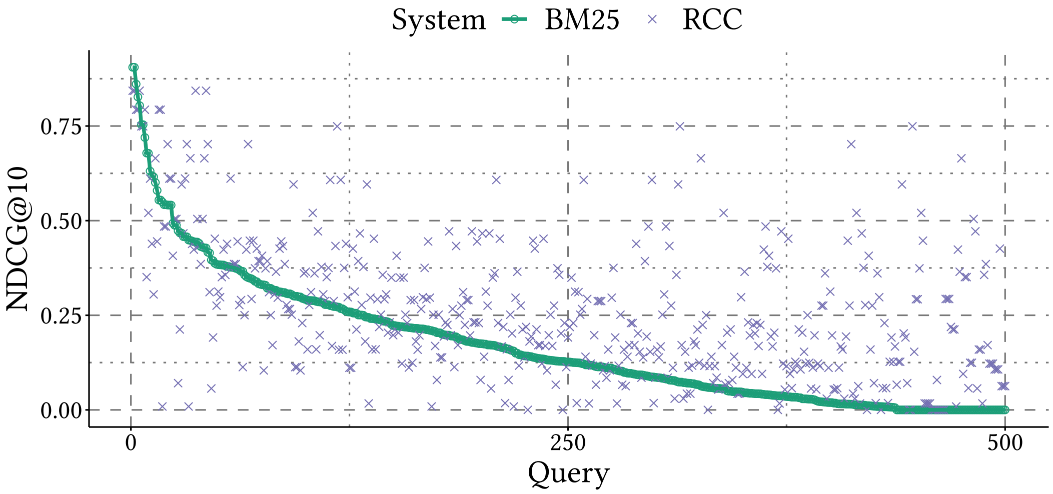


Figure 3.1: NDCG@10 effectiveness of 500 queries between two different rankers. Queries are ordered by their monotonically decreasing BM25 NDCG@10 score.

permanently altered perception that the search engine is ineffective. Song et al. [181] describe search abandonment as an issue that large-scale web search providers carefully monitor to ensure good quality of service.

To mitigate the risk of swapping to a ranker that impacts searcher satisfaction, IR evaluation researchers have considered various *overlays* on effectiveness metrics that aim to preserve the existing effectiveness of a system when considering an alternative ranker. (These risk overlays were surveyed in Section 2.5 on page 50.) One of the key innovations in this space is the URisk [207] overlay, which transforms the arithmetic mean of the differences in per-topic effectiveness metric scores between a challenger and a champion system to scale losses by a factor of α. The α parameter models different loss aversions where the traditional values reported are $\alpha \in \{1, 2, 5, 10\}$ [22, 23, 55, 56, 65, 66, 81, 95, 130, 182]. An important characteristic that Dinçer et al. [67] identified as lacking in the URisk overlay is that it requires sighting the risk values of many other systems to interpret their relative risk profiles. Dinçer et al. studentized those adjusted values to remove the need to intuit rows of URisk values, providing a risk overlay with statistical confidence.

Avoiding the need to evaluate many different rankers is of great advantage to practitioners with limited alternative systems and resources, for whom an expansive evaluation campaign may be prohibitively expensive. While including inference in risk evaluation is an important advance, the appropriateness of inferential testing on paired IR score differences has been scrutinized for decades; particularly when parametric testing is concerned. Inspired by renewed interest in this topic by Urbano et al. [197] and Parapar et al. [145], this chapter explores the appropriateness of applying parametric tests on risk-adjusted paired IR scores. It seeks to determine whether scaling one side of a score distribution may skew the distribution away from normality enough to bias the outcomes of the renowned IR parametric Student t-test, and suggests alternatives that account for this issue if it is observed to be problematic.

To explore how amenable risk inferences are to parametric testing, the problem is addressed from first principles by extending properties of existing risk measures and interpreting whether their adoption might yield practically different outcomes. Current risk overlays measure risk-reward trade-offs piece-wise, with effectiveness losses multiplied by a constant scalar. The calculus of slight drops in effectiveness is the same as substantial decreases at a per-query level. Previous work shows that searchers tend to ignore minor changes in effectiveness scores, so an interesting question is whether all deltas should be treated the same. In IR experimentation, Turpin and Scholer [193] found that the only reliable signal of whether retrieval effectiveness scores impacted task performance was whether the top document in a ranking was relevant. Allan et al. [12] saw that the relationship between bpref effectiveness and recall fits an S-shaped mathematical expression, where there is a "large intermediary region in which the utility difference is not significant". Whether modeling risk with a hypothetical smooth S-shaped value function changes the relative system risk rankings against the standard linear method for various IR risk overlays is explored in this chapter, along with how amenable the smooth risk-adjusted scores are for use with the TRisk measure.

Included in this chapter are a wide range of meta-statistics on risk-adjusted IR score distributions against various normality measures, while also exploring the stability of confidence intervals on changing risk-level parameters. Five different significance testing approaches with contrasting assumptions about the shape of the distribution are examined, extending previous work that leverages bootstrapping. Bootstrapping risk-adjusted values enables the ability to report confidence intervals and visualize the shape of the distribution, where Sakai [158] advocates reporting as much information about the significance testing approach used as possible, rather than p-values alone. As sample size also impacts how sensitive tests are and whether the central limit theorem has influenced normality, the Robust04 collection is observed to consider cases where 250 topics are available, instead of the typical collection size of 50.

Contributions. This chapter addresses the thesis research question and sub-questions:

Research Question (RQ3): *How do the distributional properties of risk-adjusted scores impact the results of parametric statistical tests?*

S-RQ3.1: *How much do the system rankings change when a hypothetical smooth value function that weighs outliers more heavily is juxtaposed with the standard linearly weighted risk value function?*

S-RQ3.2: *How do the distributional properties of URisk-adjusted scores change when the treatment of the losses is varied?*

S-RQ3.3: *How do parametric and nonparametric confidence intervals differ when performing risk-based evaluation?*

S-RQ3.4: *How does evaluating larger topic samples affect confidence intervals when performing risk-based evaluation?*

This chapter contributes the first investigation into a different risk value function on IR scores, as well as a novel exploration of the distributional properties of risk-adjusted scores and how the risk parameter influences the veracity of the inferences generated.[1] The methodology presented describes several transformations made to existing risk overlays to support more complex risk value functions. Despite the theoretical benefits of a risk value function that weighs outlier differences in effectiveness more than small ones, which Allan et al. [12] conjecture to be more likely to notice changes in search quality, the smooth cubic value function explored did not yield substantially different system orderings to the standard linear method (S-RQ3.1). In examining the distributional shape of the adjusted scores generated between value functions, which might meaningfully alter the outcomes of parametric significance tests that assume normality, the results show that the explored smooth approach

[1] This chapter combines contributions from two papers. The first is R. Benham, A. Moffat, and J. S. Culpepper. On The Pluses and Minuses of Risk. *Proc. Asia Information Retrieval Societies Conf. (AIRS)*, pages 81–93, 2019. The second paper is: R. Benham, B. Carterette, A. Moffat, and J. S. Culpepper. Taking Risks with Confidence. *Proc. Australasian Document Computing Symp. (ADCS)*, p. 1–4, 2019.

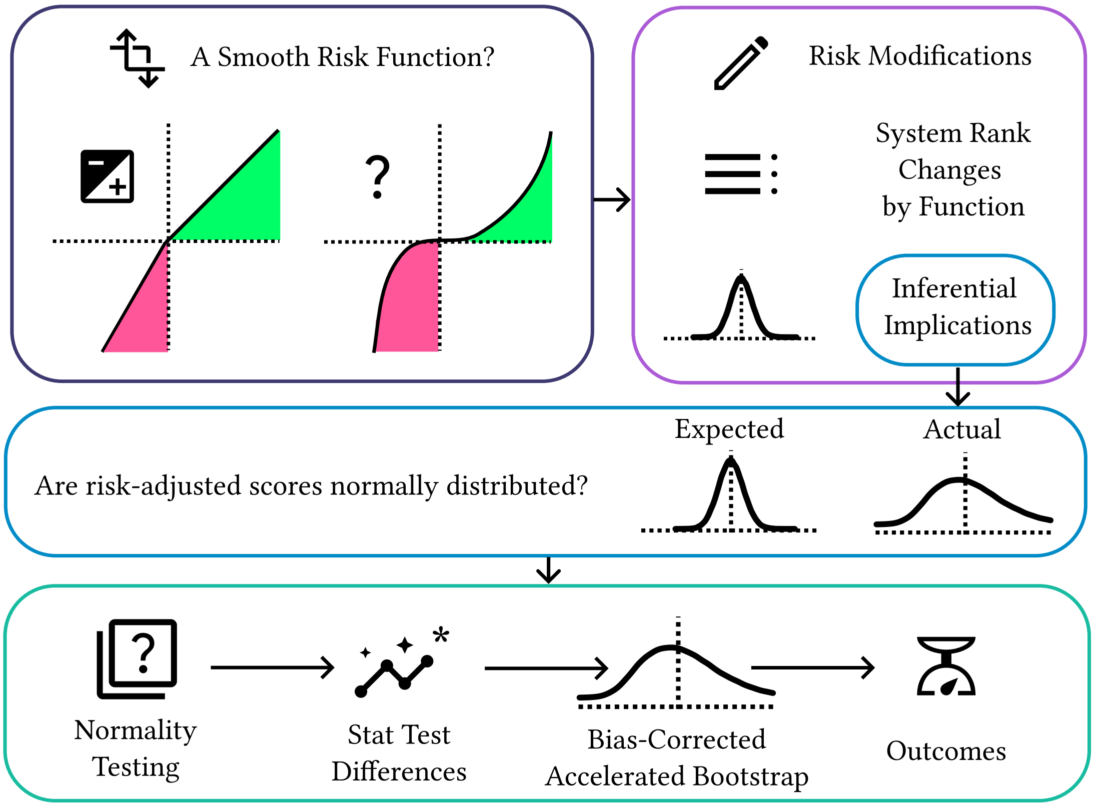


Figure 3.2: A road-map of the chapter, showing the risk concepts explored from left-to-right. A novel investigation is first presented into the usefulness of existing risk value functions against new alternatives with interesting contrasting properties. These risk-adjusted functions may have different distributional properties than typical IR effectiveness scores. That may affect inferential testing approaches, which assume that the values follow a particular shape. Finally, a new method to perform risk inference in an IR context is proposed, informed by the inferential implications of previous steps.

deviated from normality more than the linear method, but also found that risk-adjustments may influence in the shape of the distribution (S-RQ3.2). Further experimentation with different corpora and baseline runs showed that risk inferences using parametric methods appear to disagree with their nonparametric counterparts for the typically reported IR risk parameters (S-RQ3.3), finding that more topics than the typical number present in test collections are required to support consistent inferential analyses (S-RQ3.4). These results indicate that statistical tests honoring skew may be more accurate than those which assume normality on risk-adjusted scores, and culminate in a broader understanding of inferential risk overlays.

Figure 3.2 presents a road-map of the chapter. The chapter begins by taking stock of naming and formulaic issues with risk overlays in Section 3.1 commencing on page 63 and adopts modifications used throughout the thesis from therein. Section 3.2 on page 74 explores the case for a smooth risk value function that weighs outlier score differences as more important than the standard linear approach, and explores how the relative system orderings and distributional properties change using a range of meta-statistics. As bootstrapping a sample statistic can provide information about distributional properties, as well as produce different forms of confidence intervals for statistical inference, Section 3.3 starting on page 80 explores the effect that the risk-level has on parametric and nonparametric statistical testing approaches on topic sizes of either 50 or 250; on a different test collection-metric pair. The chapter ends in Section 3.4 on page 93 by answering the research questions developed through the above experiments.

3.1 Risk Functions

The risk transformation that biases the losses part of a score difference distribution is ubiquitous across many IR risk overlay functions. That risk overlay penalizes a small unnoticeable drop in effectiveness with the same calculus as a large drop in effectiveness. This section explores an alternative continuously differentiable function, juxtaposed with the standard piece-wise function. These experiments advance knowledge towards an improved risk evaluation methodology when deliberating whether to swap a deployed ranker with an alternative model.

The section begins by examining aspects of the risk overlay equations that have room for improvement: the α loss aversion parameter and the wording of risk measures against the direction of the vector quantities output. After proposing solutions to ambiguous aspects of the standard risk transformation applied in IR, a set of experiments for comparing the behavior of an alternative *smooth* risk value function are defined, which seek to weigh outliers more heavily than the standard approach. Of interest is whether smooth value functions give substantially different system risk profiles compared to their linear counterparts. Outcomes in this section address S-RQ3.1: *How much do the system rankings change when a hypothetical smooth value function that weighs outliers more heavily is juxtaposed with the standard linearly weighted risk value function?*

3.1.1 A Revised Standard Risk Function

Recall from Equation (2.39) on page 53 that the URisk equation is:

$$URisk_\alpha = \frac{1}{|T|}\left[\sum_{j\in T^+}\Delta_j - (1+\alpha)\cdot\sum_{j\in T^-}\Delta_j\right] , \tag{3.1}$$

where T^+ represents the set of topics where the challenger system outperforms the champion system, and $(1+\alpha)$ scales the impact of opposing effectiveness score differences T^- for a given pair of systems. The α parameter linearly scales losses relative to the champion for a researcher-selected loss aversion.

The α Trade-Off Parameter. For each paired score difference between a champion system and a challenger, the α parameter linearly scales losses relative to the champion for a researcher-selected loss aversion. Table 3.1 on the next page lists the α parameters selected in a sample of ten research papers that used a risk overlay as part of their analysis. From

α	Citations
2	Gallagher et al. [81] [b], Benham et al. [23] [b]
5	Collins-Thompson et al. [56] [a], McCreadie et al. [130] [a], Yılmazel and Arslan [218] [b]
1, 2, 3, 4	Hashemi and Kamps [95] [a]
1, 5	Manotumruksa et al. [127] [a, b]
1, 5, 10	Collins-Thompson et al. [55] [a], Dinçer et al. [66] [a], Sousa et al. [182] [a, b], Benham and Culpepper [22] [a, b]
2, 5, 10	Rodrigues et al. [153] [d]
1, 5, 10, 20	Dinçer et al. [65] [a, b, c, d]

[a] URisk [b] TRisk [c] ZRisk [d] GeoRisk

Table 3.1: Differing sets of α employed for risk evaluation in other studies.

this sample, the values practitioners use are within the range $\alpha \in \{1, 2, 5, 10\}$, with some marginal variation irrespective of the risk overlay used. If a practitioner explored a risk setting counting losses twice as much as gains, they may set $\alpha = 1$ and evaluate the results. That "twice as much" objective could introduce "off by one" [119] errors when $\alpha = 2$. That is because risk overlays in their current form treat $\alpha = 2$ losses with a *three*-fold penalty.

Naming and Signed Direction Of "Risk" Values. Suppose that an experiment has been conducted between a challenger and a champion system, and a practitioner wishes to perform a risk evaluation. Ideally, the challenger system will exhibit low risk for some loss aversion parameter α. If the calculated value for a risk overlay is in the positive direction, then the reward of each per-topic effectiveness score difference outpaces the effectiveness losses for α. The current interpretation of URisk values are that large numbers are preferable, but when communicating those outcomes, the phrasing of the "risk" result does not match the direction of the risk vector. The practitioner wants to maximize their "risk" number, however when discussing it, they often must reexplain that large risk values indicate reward, not risk.

Solutions. Solutions to the above formulaic issues identified are now proposed, reforming the existing measures to improve the clarity of reported risk results.

To address the directional issues of risk overlays, URisk defined in Equation (2.39) is redefined as URisk^{-}, with definition:

$$URisk_r^- = -URisk_{r-1} = \frac{-1}{|T|} \left[\sum_{j \in T^+} \Delta_j - r \cdot \sum_{j \in T^-} \Delta_j \right] . \tag{3.2}$$

Collection	Citation	Documents	Unique Terms	Total Terms	Topics
ClueWeb09A	[1]	503,903,810	511,580,198	343,756,256,409	50

Table 3.2: Statistics for the ClueWeb09A collection used to explore the distributional properties of smooth and linear risk value functions.

Additionally, TRisk specified by providing URisk adjusted values to the Student t-value function in Equation (2.23) is redefined to TRisk^{-}:

$$TRisk_r^{-} = -TRisk_{r-1} = \frac{URisk_r^{-}}{se(URisk_r^{-})} = \frac{\overline{-URisk_r^{-}}}{sd(URisk_r^{-})/\sqrt{|URisk_r^{-}|}}\,, \tag{3.3}$$

where *se* is the standard error of the URisk^{-} adjusted scores. Finally the ZRisk and GeoRisk overlays defined in Equation (2.40) and Equation (2.43) adopt the new forms:

$$ZRisk_r^{-}(S_i) = -ZRisk_{r-1}(S_i)\,; \tag{3.4}$$

$$GeoRisk_r^{-}(S_i) = -GeoRisk_{r-1}(S_i)\,. \tag{3.5}$$

The directions of the risk vectors have been reversed in these revised versions, and the $(\alpha + 1)$ component has been replaced with r to distinguish it from α, with $r = 1 + \alpha$. These revised definitions are used in the context of work conducted within this thesis, and r no longer refers to a relevance vector of a ranking as it had in Section 2.2.

3.1.2 Smooth Value Function Methodology

Using the above definitions for the standard risk value function, this section now focuses on a novel investigation into the potential usefulness of a smooth value function used in a risk overlay. A hypothetical S-shaped polynomial intuited from searcher task performance of changes in IR effectiveness from the S-shaped observation in Allan et al. [12] is explored, using the interquartile boundaries of outlier score differences concerning a baseline system (using the Tukey [190] definition of an outlier, $1.5 \times IQR$). Based on this S-shaped principle, a hypothesis is explored where if disagreement in relative risk between the standard approach is prominent, that the smooth value function might yield answers that better reflect the attitudes of searchers. To model that goal, an initial investigation is restricted to a collection with an associated set of runs, a baseline, and an appropriate metric. Note that the fit model does not aim to generalize to other corpora/run/metric combinations and that the purpose of this investigation is to provide early work in understanding the statistical properties of risk-adjusted scores. As seen in Table 3.1 on the previous page in Section 3.1.1, the risk-level observed in many studies are biased towards lower values, with $\alpha = 5$ being a central reference point. For a conservative exploration of risk values on smooth risk functions, $r = 2$ is used to count losses twice as much as gains.

Experimental Setup. Since defining the smooth value function depends on knowing the regions of score differences that constitute outliers, the Dinçer et al. [65] evaluation scenario used to motivate ZRisk and GeoRisk is used as an exemplar. The collection applied is the 2012 TREC Web Track [48] where the ClueWeb09A[2] corpus was used. In the collection, 48 runs were submitted to the track and assessed for relevance against 50 topics, and ERR@20 was the official metric. Table 3.2 on the previous page shows summary statistics about the ClueWeb09A corpus.

Later, the 2013 TREC Web Track [55] used the ClueWeb12[3] corpus and considered URisk against an Indri baseline with a query likelihood ranker and spam filtering applied [66]. The track organizers also distributed that same baseline for use on the 2012 TREC Web Track on the ClueWeb09A corpus, which Dinçer et al. [65] calls `indriCASP`; this chapter continues to use this naming for consistency.[4] Also adopted from the methodology of Dinçer et al. [65] and Collins-Thompson et al. [55] are the risk comparisons against the most effective submitted run per research group, a combined set of eight runs. The ERR@20[5] score of `indriCASP` is 0.195, and the median ERR@20 score of the nine systems (top-eight runs combined with the `indriCASP` baseline) is 0.220 (matching the ERR@20 score of `utw2012c1`), so `indriCASP` is competitive among this set of challenger systems.

A hypothetical smooth value function that handles a region of scores within typical variation differently to significant score differences (outliers) is derived using a range of points to form a cubic regression. A minimum of four points are required for a cubic regression, where more points can be used when curve fitting techniques are applied (such as, the `lm` function distributed with the `R` programming language). Three Cartesian pairs corresponding to known aims of the smooth function in regard to extreme score differences are initially selected. Recall that ERR@20 score differences are bounded in the range $[-1, 1]$. The first of the points (best thought as vector pairs of original value and risk-transformed value, they are displayed shortly in Figure 3.4 on page 68) used to define the risk cost function $(1, 1)$ corresponds to the standard URisk mapping in the most extreme case, where gains are not transformed with respect to the magnitude. As $r = 2$ is set to count losses twice as much as gains, the second point $(-1, -2)$ is the opposite extreme with the risk transformation applied. The last semantically valuable mapping is $(0, 0)$, where an equal score difference maps to no change in risk. To ascribe greater importance to outlier score differences using the smooth value function, two additional points are included when finding the cubic of best fit using least summed square errors. These points correspond to the $1.5 \times \mathit{IQR}$ boundaries when pooling every score difference between the `indriCASP` baseline and all 48 runs, where the losses-side of the equation is doubled and the gains-side remains unchanged.

[2] ClueWeb09A dataset: `http://lemurproject.org/clueweb09/`

[3] ClueWeb12 dataset: `http://lemurproject.org/clueweb12/`

[4] Run data available at: `https://github.com/trec-web/trec-web-2013`

[5] Evaluated using `gdeval` from TREC 2013 Web Track repo above.

Finding The Outlier Points. The smooth risk value function has been designed to weigh outliers more heavily. To that end, the dispersion of score differences of `indriCASP` on the ERR@20 metric on different systems is observed. Figure 3.3 on the next page shows box plots of the ERR scores for top-performing runs, as well as the pooled score differences on all 48 reference systems. In observing the number of outliers present on each system, there is a possibility that an S-shaped function may order runs differently compared to a linear risk function. Upon inspecting the "All" label right-most in the plot, the fences of the box plot are present at -0.241 and 0.292. These bounds are used to form the remaining points in the cubic regression. Again, note that these outlier figures are tied to the ERR@20 metric, the TREC Web Track collection, and the 48 systems against the `indriCASP` baseline; the $s(\Delta)$ function is not intended to generalize to other collections.

Function Definitions. Recall above, three points that semantically map to the objective of the explored smooth value function are selected: $(-1,-2),(0,0),(1,1)$. As the fences of the box plot for score differences between `indriCASP` and all 48 other reference systems were -0.241 and 0.292, these ranges are now mapped into point form: $(-0.241,-0.05)$, and $(0.292, 0.05)$. The value ± 0.05 was selected by trial-and-error, ensuring that the resulting cubic of best fit is one-to-one to allow consistent rankings to be produced (discussed below). Changes within the ± 0.05 region are conjectured to be of less psychological value than differences outside of this region. Since a searcher is less likely to notice slight score differences [12, 193], the extreme differences (in risk and reward) should ideally count for more.

With the points of interest for the risk goals identified, the `R lm` function is used to compute a cubic regression across these pairs, setting the y-intercept term $d = 0$ from the resultant cubic of best fit result to ensure intersection with the origin point $(0,0)$. The resultant smooth function takes the form:

$$s(\Delta) = 1.38426\Delta^3 - 0.51659\Delta^2 + 0.11578\Delta\,, \tag{3.6}$$

where $\Delta \in [-1,1]$ corresponds to a score difference in ERR@20 against `indriCASP` on the TREC web track collection. The standard linear piece-wise function in the URisk$^-$ family of measures has the definition:

$$l(\Delta) = \begin{cases} \Delta & \Delta \geq 0 \\ r \cdot \Delta & \Delta < 0\,, \end{cases} \tag{3.7}$$

where $\Delta \in [-1,1]$ is the difference in effectiveness between baseline(s) and a run, and r corresponds to the risk-level parameter. The $r = 2$ risk level is set to compare the outcomes of $s(\Delta)$ and $l(\Delta)$. Figure 3.4 on the next page plots the smooth and linear value functions against each other between the limits of possible ERR@20 score differences in the range $[-1,1]$, where crosses mark the points used to form the cubic of best fit.

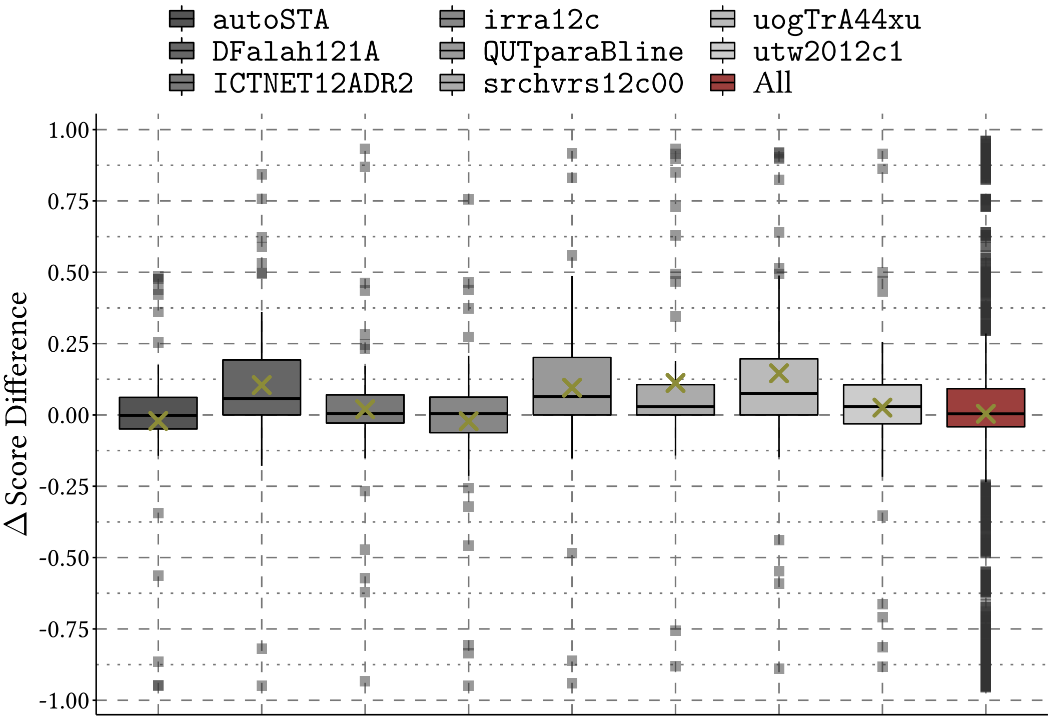


Figure 3.3: Differences in ERR@20 for eight systems relative to the `indriCASP` baseline, for the TREC 2012 Web Track corpus. The "All" box plot shows the score differences against all submitted runs to the track; crosses indicate arithmetic means.

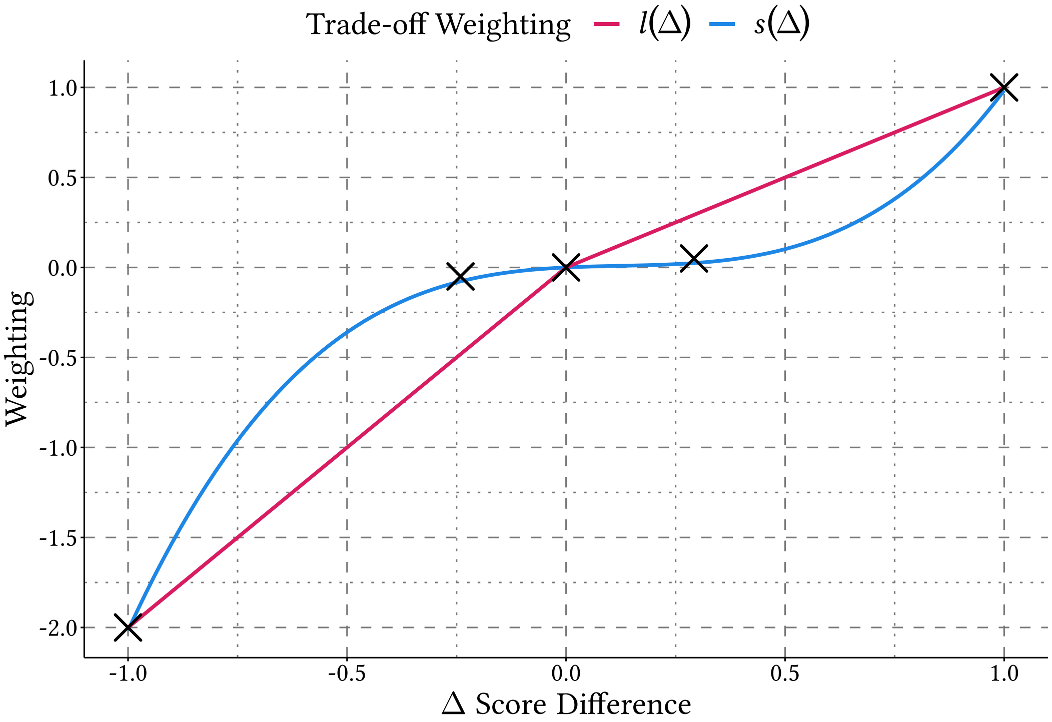


Figure 3.4: The linear URisk$^-$ function $l(\Delta)$ with the parameter $r = 2$, versus the explored cubic regression variant, $s(\Delta)$. Crosses mark the points used to model $s(\Delta)$.

On Smooth Cost Function Consistency. Wang et al. [207] explored the benefit of URisk as an objective function in learning-to-rank (LtR), where Wang et al. demonstrated that the standard form of URisk has the property of being *consistent*. This notable property is also present in the smooth value cost function, $s(\Delta)$. Although no LtR experiments are run in this thesis, there exists a case where $s(\Delta)$ can be applied in place of $l(\Delta)$ in an LtR setting.

To demonstrate that $s(\Delta)$ produces consistent results, the derivative of $s(\Delta)$ is $s'(\Delta) = 4.15278\Delta^2 - 1.03318\Delta + 0.11578$; found using the polynomial differentiation rule on each of the terms of $s(\Delta)$, where $f(x) = x^n$, $f'(x) = nx^{n-1}$. Since the discriminant of $s'(\Delta)$ is = -0.005 (the square root part of the quadratic formula, $b^2 - 4ac$, as $s'(\Delta)$ is quadratic), it has no real solutions, meaning $s(\Delta)$ must be one-to-one. Moreover, since $s'(\Delta)$ is of the form $a\Delta^2 + b\Delta + c$ where $a > 0$, $s'(\Delta)$ only returns positive values. With that in mind, knowing that $s(\Delta)$ is one-to-one, $s(\Delta)$ is strictly monotonically increasing. It follows that score differences Δ_i and Δ_j cannot be swapped inconsistently with $s(\Delta)$, provided that the evaluation metric also has the property of being consistent. Since $s(\Delta)$ is consistent, it may be used as a loss function in learning to rank instead of the typical $l(\Delta)$ option.

3.1.3 Transforming Existing Risk Overlays

Section 3.1.2 explored the case for a smooth risk value function and demonstrated useful properties in an IR context. This subsection contrasts the examined smooth $s(\Delta)$ loss function with the the original $l(\Delta)$ to observe whether different risk overlays on each function produce different relative system risk rankings. As URisk$^-$ returns the mean of the $l(\Delta)$ values, it can be replaced with $s(\Delta)$ instead, with the average of those values calculated. Similarly, TRisk$^-$ studentizes $l(\Delta)$ values, allowing $s(\Delta)$ values to be studentized instead. How amenable the $s(\Delta)$ adjusted scores are in a parametric inferential context is investigated later in Section 3.2.1.

Finding that ZRisk$^-$ produces the same result if the risk transformation is applied *before* standardization, $s(\Delta)$ can be evaluated without renormalizing multiple times, denoted RiskZ$^-$. As a reminder, ZRisk$^-$ standardizes the scores before the trade-off value function is applied. As RiskZ$^-$ and ZRisk$^-$ calculations provide equivalent answers, GeoRisk$^-$ can be evaluated straightforwardly using a smooth value function, as it combines the ZRisk$^-$ result with the (arithmetic) mean system effectiveness.

Consider the ZRisk$^-$ equation:

$$\begin{aligned} \mathit{ZRisk}_r^-(S_i) &= -1 \cdot \left[z_i^+ + r \cdot z_i^-\right] \\ &= -1 \cdot \left[\sum_{j \in Q^+} \frac{S_{ij} - e_{ij}}{\sqrt{e_{ij}}} + r \cdot \sum_{j \in Q^-} \frac{S_{ij} - e_{ij}}{\sqrt{e_{ij}}}\right] \\ &= -1 \cdot \left[\Delta_z^+ + \Delta_z^-\right] \ , \end{aligned} \tag{3.8}$$

where Q^+ represents the instances where the score difference from the expected value $S_{ij} - e_{ij}$ is positive, and Q^- where the score difference is negative. Both summation terms are redefined to Δ_z^+ and Δ_z^- to aid in showing that RiskZ$^-$ yields equal results to ZRisk$^-$. Based on this, RiskZ$^-$ can be defined as:

$$\begin{aligned} \mathit{RiskZ}_r^-(S_i) &= -1 \cdot \left[\sum_{j \in T^+} z(S_{ij} - e_{ij}) + \sum_{j \in T^-} z(r \cdot (S_{ij} - e_{ij})) \right] \\ &= -1 \cdot \left[\Delta_*^+ + \Delta_*^- \right] . \end{aligned} \tag{3.9}$$

Focusing on the Δ_*^+ part of RiskZ$^-$, the standardized score from the standard normal distribution can be evaluated as:

$$\Delta_*^+ = \sum_{j \in T^+} \frac{S_{ij} - e_{ij} - E\left[S_{ij} - e_{ij}\right]}{\sqrt{e_{ij}}} . \tag{3.10}$$

To evaluate the expected value $E\left[S_{ij} - e_{ij}\right]$, observe that the expectation operator $E\left[\cdot\right]$ is linear. Hence, $E\left[S_{ij} - e_{ij}\right] = E\left[S_{ij}\right] - E\left[e_{ij}\right] = 0$, and therefore for the reward component,

$$\Delta_*^+ = \sum_{j \in T^+} \frac{S_{ij} - e_{ij}}{\sqrt{e_{ij}}} = \Delta_z^+ . \tag{3.11}$$

On the risk-adjusted component, $E\left[cX\right] = c \cdot E\left[X\right]$, meaning that a similar argument allows:

$$\Delta_*^- = r \cdot \sum_{j \in T} \frac{S_{ij} - e_{ij}}{\sqrt{e_{ij}}} = \Delta_z^- . \tag{3.12}$$

That is, $\mathit{ZRisk}_r^-(S_i) = \mathit{RiskZ}_r^-(S_i)$. Finally, since calculating risk before normalizing gives the same result as ZRisk$^-$, the smooth function s is used in place of the existing linear scaling applied in ZRisk$^-$ to form the $s(\Delta)$ variant:

$$\begin{aligned} &= -1 \cdot \left[\sum_{j \in T^+} z(s(S_{ij} - e_{ij})) + \sum_{j \in T^-} z(s(S_{ij} - e_{ij})) \right] , \\ &= -\sum_{j \in T} z(s(S_{ij} - e_{ij})) . \end{aligned} \tag{3.13}$$

3.1.4 System Rank Agreement

After defining modified versions of risk overlays to suit a smooth and linear risk value function, this subsection computes the relative changes in system orderings between the two approaches to answer S-RQ3.1. Table 3.3 on the following page shows the difference in weighting functions across the top-eight systems from the 2012 Web Track across each of the considered risk overlays. The highlighted cells indicate the least risky system on each risk value function for different risk overlays. Interestingly for the URisk$^-$ overlay, noting this is a small

	URisk$^-$		TRisk$^-$		ZRisk$^-$		GeoRisk$^-$	
System	$l(\Delta)$	$s(\Delta)$	$l(\Delta)$	$s(\Delta)$	$l(\Delta)$	$s(\Delta)$	$l(\Delta)$	$s(\Delta)$
`autoSTA`	0.12	0.10	1.63	1.85	8.14	0.91	−0.31	−0.30
`DFalah121A`	−0.05	0.02	−0.85	0.52	7.02	0.61	−0.41	−0.39
`ICTNET12ADR2`	0.05	0.03	0.78	0.75	6.80	0.17	−0.35	−0.33
`irra12c`	0.12	0.08	1.70	1.72	5.86	0.27	−0.31	−0.29
`QUTparaBline`	−0.04	0.03	−0.57	0.62	6.51	0.97	−0.40	−0.38
`srchvrs12c00`	−0.07	−0.02	−1.08	−0.50	8.53	1.42	−0.42	−0.40
`uogTrA44xu`	−0.09	−0.03	−1.29	−0.63	5.29	0.70	−0.43	−0.41
`utw2012c1`	0.06	0.05	0.79	1.15	6.33	0.37	−0.35	−0.33

Table 3.3: Risk-adjusted scores for the linear $l(\Delta)$ and smooth $s(\Delta)$ variants of risk value functions on several popular risk overlays, for the top-eight runs submitted to the 2012 TREC Web Track, scored using the ERR@20 metric. URisk$^-$ and TRisk$^-$ have their risk scores calculated against the `indriCASP` champion, where ZRisk$^-$ and GeoRisk$^-$ scores are relative to all submitted systems to the track (including `indriCASP`). Highlighted values indicate the least risky system in that column among the eight considered. No TRisk$^-$ values are highlighted as these values are in units of URisk$^-$ per standard error of URisk$^-$, where standard error is related to only two systems.

sample size of runs, that when the linear function $l(\Delta)$ reports a high risk, the relative change in URisk$^-$ score between the $s(\Delta)$ function is negligible. However, when $l(\Delta)$ describes that a system is more rewarding, the corresponding $s(\Delta)$ risk adjustment is greater in magnitude. Ranking TRisk$^-$ t-values by different system outputs is invalid, as these values are expressed the units URisk$^-$ per standard error of URisk$^-$, where the standard error relates to two systems only. Therefore, these TRisk$^-$ values are not highlighted in the table. The `uogTrA44xu` system is the most rewarding system according to both $s(\Delta)$ and $l(\Delta)$ on all risk overlays except for ZRisk$^-$, where the $s(\Delta)$ method rewards the `ICTNET12ADR2` run instead. For the TRisk$^-$ overlay, no values fall outside the $-2.0 < t < 2.0$ statistical boundaries in the linear or smooth functions.

As can be seen from Table 3.3, the linear value function appears to have strong agreement with its smooth counterpart when comparing risk values, suggesting that they do not generate different outcomes. Further, while the relative system rankings of ZRisk$^-$ are different between the two value functions, when integrated into GeoRisk$^-$ their system risk rankings are equal. It is critical to recognize that this is one study on one collection, but these interim results suggest that the experimental risk value function explored correlates well with the traditional linear function. To fully answer S-RQ3.1, the next paragraph applies the rank-biased overlap similarity metric to all reference systems to further quantify the difference between the two risk value functions.

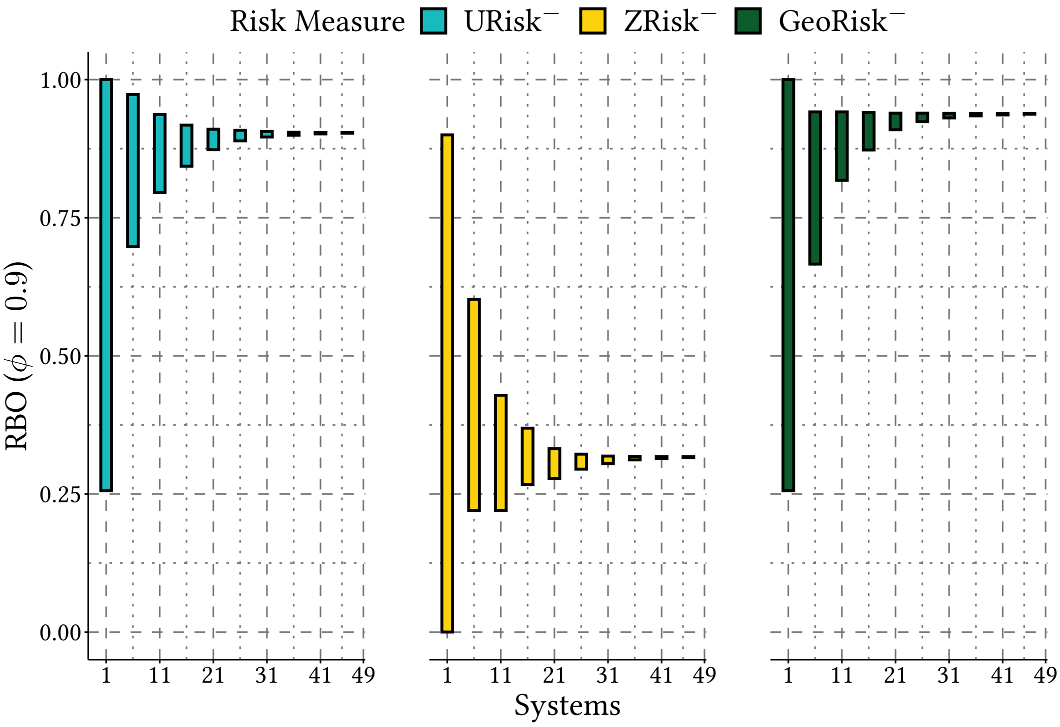


Figure 3.5: RBO ($\phi = 0.9$) rank similarity scores when an increasing set of systems (ordered by ERR@20 effectiveness) is compared between the linear $l(\Delta)$ and smooth $s(\Delta)$ variants over risk overlays amenable to ranking. A concrete example of comparing the relative ranking of eight systems is available in Table 3.3 on the previous page, where the goal of this plot is to systematize comparing many group sizes. The risk-level for each overlay is set at $r = 2$, where URisk⁻ uses `indriCASP` as the champion, and ZRisk⁻ and GeoRisk⁻ correspond to all reference systems including `indriCASP`.

System-Risk Rank Similarity. Ideally, the smooth risk value function would give meaningfully different results against the simpler linear approach to justify the overhead in specifying the function. The previous discussion demonstrates that the system-risk rank similarity between $l(\Delta)$ and $s(\Delta)$ functions are very similar based on the eight-most effective submitted runs for each group, but it is by no means exhaustive. To fully answer research question S-RQ3.1, the similarity of the ranked systems from least-to-most risky are between the two risk transformation approaches is now explored. All systems in the collection are considered, with each of $l(\Delta)$ and $s(\Delta)$ implemented for comparison in URisk⁻, ZRisk⁻, and GeoRisk⁻. As varying the suffix lengths of the count of systems compared between each risk function produces a non-conjoint ranking (or, an incomplete ranking), similarity measures such as Kendall's τ or Spearman's ρ will give rise to invalid values in that case since they assume a full ranking has been observed. (There is an infinite number of systems that could have been compared against, as the relevance judgments are able to compute evaluation scores for more than the pooled systems.) The *rank-biased overlap* (RBO) measure proposed by Webber et al. [210] is used since it meets the non-conjointness requirements and it can generate top-weighted similarity scores. The top-heaviness is a useful property of RBO in this context, as the outcomes of this experiment are more concerned with changes about the least risky systems; perturbations at the bottom of the system order are less important. The persistence parameter $\phi = 0.9$ is fixed for all experimentation; an RBO value of 1.0 indicates complete agreement, and 0.0 indicates no agreement.

The `indriCASP` run is used as a baseline for risk adjustment purposes against the total set of 48 reference systems. To measure the similarity of the system orders produced using different risk value functions, the number of systems is gradually increased between 1 and the total count of 49 (48 for URisk^{-} excluding `indriCASP`) in steps of 5, with their RBO computed. Each are included in the ranking comparison in descending ERR@20 effectiveness to visualize the convergence towards the complete rank similarity value on the 49 systems.

Figure 3.5 on the previous page shows the resultant RBO ($\phi = 0.9$) values. As the expected viewing depth of the similarity measure is $1/(1-0.9) = 10$ ranked values, the distance between the minimum overlap score and maximum for the 1 and 6 system is understandably quite large. The lower value of each bar is the minimum RBO value for that set size, and the top-value is the maximum possible value, the sum of the lower bound and the RBO residual. As Webber et al. [210] show the prefix of a pair of lists increases in size, "the range of the possible full similarity score decreases monotonically"; this convergence is prevalent for all risk overlays compared. When all reference systems are considered, only ZRisk^{-} appears to be ranking systems differently with the smooth value function, with a score converging on 0.316. But, when that different value function is used in combination with GeoRisk^{-} (which is important as ZRisk^{-} strips the effectiveness magnitude out from each system), it is clear that including the effectiveness back into the value is dominating the influence of the ranking with a compelling similarity of 0.937 and a negligible residual. URisk^{-} has a marginally smaller similarity score of 0.903, where the lower bound converges rapidly with a hyperbolic shape on both risk overlays. These results provide an answer to S-RQ3.1, as the smooth value function ranks systems similar to the standard value function, except for ZRisk^{-}.

3.1.5 Summary

This section has explored the potential of using a smooth weighting function on IR risk evaluation overlays, which weighs gross score differences more than meager ones with a variable rate of change. That intuition is modeled on the assumption that searchers notice appreciable differences in effectiveness more than small ones [12, 193], and was important to investigate as it is not an explicit goal of existing risk measures. Several popular risk measures were adjusted to explore the difference between an experimental smooth risk value function against the traditional linear piece-wise approach. In experiments using the ERR metric in line with previous studies and TREC 2012 Web Track data, no evidence was found that would indicate that using a smooth risk function might lead to substantially different evaluation outcomes when undertaking a risk-sensitive experimental comparison; answering S-RQ3.1. Although this investigation showed that the smooth and traditional risk-adjusted scores generally agree (with the exception of ZRisk^{-}), it is possible that the distributional properties of the smooth risk-adjusted scores could be more amenable for statistical testing. Overall, the close similarity between the methods is a net positive for the traditional overlays in use in the literature. The existing piece-wise risk measures tend to implicitly agree with alternative risk measures tuned to weigh outlier score differences more heavily, which are liable to be more noticeable to searchers [12].

3.2 Distributional Properties

Section 3.1 explored the usefulness of a smooth risk value function on paired IR effectiveness score differences, breaking from the tradition of using the conventional piece-wise linear approach. Although the smooth value function yielded system risk orderings that were similar to the linear approach, the goal of this chapter is to explore how risk overlays can be extended to better support statistical decision making overall. The Student t-test used in TRisk is parametric, meaning that the test assumes a distributional property holds and then builds confidence intervals around those assumptions. As the mean distribution of smooth risk-adjusted paired values may be amenable to normality assumptions (as the linear approach is believed to be); this section examines that prospect to answer S-RQ3.2. Moreover, it contributes novel analysis in how applicable the standard linear risk-adjusted score distributions are to the assumption of normality, which has implications on the ideal kinds of statistical tests for risk inference discussed in Section 3.3 on page 80.

3.2.1 Smooth and Linear Bootstrap Distributions

> "The shape of the bootstrap distribution approximates the shape of the sampling distribution, so we can use the bootstrap distribution to check the normality of the sampling distribution."
>
> — Hesterberg et al. [100, p. 18–16]

Q-Q plots are a prevailing statistical technique used to explore how well a data sample fits a given statistical distribution; particularly when bootstrapped replicates are used. To see how well the distributions of mean risk-adjusted scores fit a Student t-distribution, using the $l(\Delta)$ and $s(\Delta)$ methods and comparing these against no transformation, the `indriCASP` system continues to be set as the champion in an initial exploration. To ensure that a challenger with an uncharacteristically high score difference is not selected, the median scoring run by ERR@20 is selected (`ICTNET12ADR3`) to perform a side-by-side comparison of each risk adjustment method. The code used to generate the bootstrap replicates was adapted from the Urbano et al. [197] investigation into different types of statistical errors produced using contrasting inferential techniques.

Figure 3.6 on the following page shows a Q-Q plot of 10,000 bootstrapped replicates of the mean URisk$^-$ values where the theoretical distribution is a `Student-t` distribution with $50 - 1 = 49$ degrees of freedom. The $s(\Delta)$ score distribution has a slightly more divergent right-tail compared to the $l(\Delta)$, which is a possible issue for t-test inferences since it

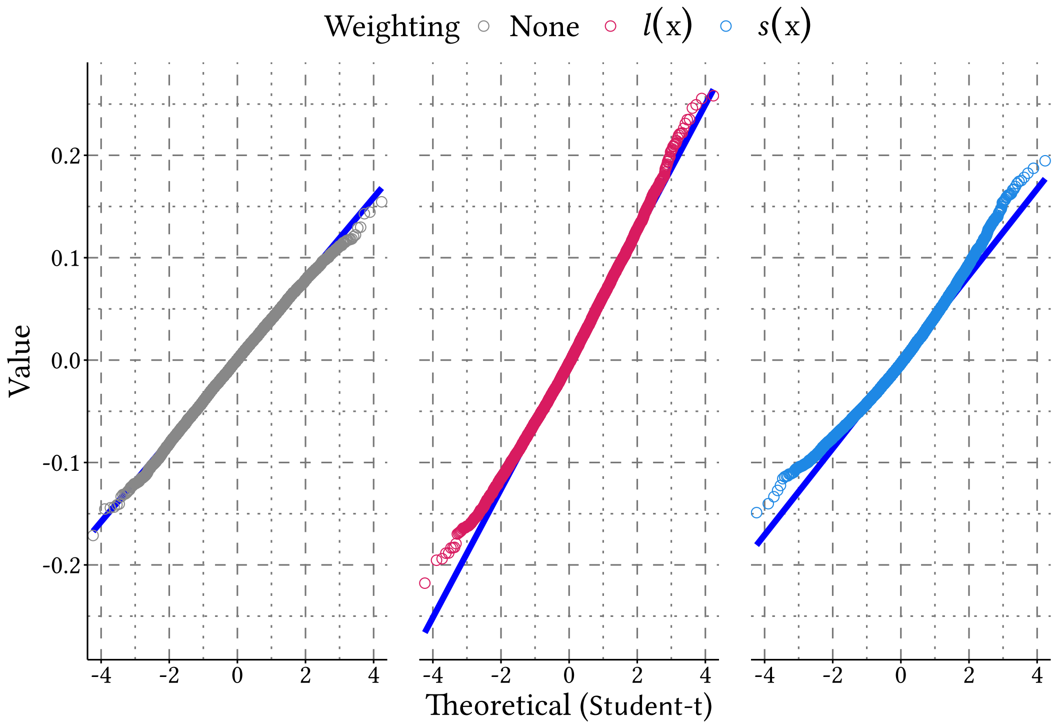


Figure 3.6: The Q-Q plot of the median scoring ICTNET12ADR3 challenger run against a indriCASP champion system. The theoretical distribution is the Student-t with $\nu = 49$ degrees of freedom. The ERR@20 metric is used, where the mean risk-adjusted score over each topic has been bootstrapped on the URisk$^-$ overlay. These replicates of the mean are compared using no risk weighting, the standard linear risk function $l(\Delta)$ with $r = 2$, and the smooth weighting function $s(\Delta)$ in Equation (3.6).

shows the bootstrap distribution of the mean may deviate from the Student distribution. That could be because sizable differences in scores correspond to a more pronounced mapping of the "risk" of changing to the ICTNET12ADR3 system. In contrast, the values from the $l(\Delta)$ and no risk functions fall along their respective reference lines, providing support for the inferences made by TRisk$^-$ using these functions on the 2012 TREC Web Track dataset. That support, however, is predicated on using $r = 2$; this experiment shows in the tail of $s(\Delta)$ (and marginally for $l(\Delta)$) that risk-adjusted scores impact how amenable the bootstrapped means of the risk-adjusted scores are to the Student distribution, and subsequently their statistical testing outcomes.

Figure 3.6 revealed that the type of risk adjustment applied on IR effectiveness scores can affect the correspondence of the bootstrapped mean to the Student distribution. Although Q-Q plots are a defensible method of assessing normality graphically, quantitative measures of normality exist which allow for broadening the scope of the analysis to many paired system comparisons. (An important caveat to this, however, is that the normality of risk-adjusted scores is not a necessary condition for the validity of the t-test, but if these scores are normal, then the means of the scores are also normal.) A commonly used gauge to test deviation from normality is the Shapiro-Wilk test [173], which produces a vector including the test statistic

Algorithm 3.1: Normality testing procedure of risk-adjusted IR effectiveness scores with leave-one-group-out.

```
Input: A set of IR effectiveness score tuples for $S_{ijk}$, where $i$ is the $i^{\text{th}}$ system, $j$ is
       the $j^{\text{th}}$ topic, and the system $i$ is contained in the $k^{\text{th}}$ group; the $P$ number of
       randomly-selected pairs of systems to compute, and the $\mathcal{R}$ set of risk-levels
       to explore.
Output: The outcomes of each of the four normality measures.
results $\leftarrow \{\}$
for $p \in P$ do
    $s_1 \leftarrow S_{random(I) \in I}$                // Randomly select a system from $I$ system indices
    $I' \leftarrow S.I \setminus (S_{s_1.k \in K}).I$         // Remove systems from same group from $I'$
    $s_2 \leftarrow S_{random(I') \in I'}$               // Random second system with constraint
    $\Delta \leftarrow s_1 - s_2$                        // Array of paired differences in scores
    $\Delta \leftarrow -1 \cdot \Delta$                      // Swap direction to refer to risk
    for $r \in \mathcal{R}$ do
        $\Delta_r \leftarrow \{\Delta^+ \cdot r\} \cup \{\Delta^-\}$          // Perform risk adjustment
                                   // Compute normality tests on risk-adjusted scores
        $W \leftarrow shapiro\_wilk(\Delta_r)$
        $K \leftarrow kolmogorov\_smirnoff(\Delta_r, \Phi, \overline{\Delta_r}, sd(\Delta_r))$
        $skew \leftarrow skewness(\Delta_r)$
        $kurt \leftarrow kurtosis(\Delta_r)$
                                                              // Append results
        $results.push(\{(p, r, skew, kurt, W.statistic, W.pvalue, K.statistic, K.pvalue)\})$
    end
end
return results
```

W and the associated p-value. The closer the W value is to one, the stronger the evidence for normality, and p-values below 0.05 pass a 95% significance test indicating deviation from normality. Results of applying this test on each kind of paired score difference distribution: no risk, $l(\Delta)$, and $s(\Delta)$; yield W test statistic values of 0.793, 0.737, and 0.460 respectively. All score difference approaches indicate a deviation from normality ($p < 0.001$). Note, however, in the case of no-risk adjustments IR practitioners are usually satisfied with the mean difference distributions being approximately normal. As the $s(\Delta)$ transformation causes an unfavorable deviation from normality for a statistical test that weakly assumes its presence while yielding similar answers to the $l(\Delta)$ transformation, we partially answer S-RQ3.2 and explore more holistically how larger values of r impact normality on the $l(\Delta)$ risk function.

3.2.2 Distributional Shape Of Risk-Adjusted Scores

In the above analysis, although there was no compelling evidence to suggest that $l(\Delta)$ when $r = 2$ results in compromised inferences, the Q-Q plot suggested that risk transformations may cause deviation from normality in the mean when losses are counted as more. The hypothesis explored now is that the severity of this issue could become more apparent on $l(\Delta)$ when r increases, noting that $r = 2$ is on the lower end of values practitioners commonly report. The previous analysis also only observed a pair of systems, where quantitative measures can be used to explore the behavior of risk-adjusted scores involving many paired system comparisons. In addition to the commonly used Shapiro-Wilk test for normality, alternative diagnostics for deviation from normality exist, such as the Kolmogorov-Smirnov test [114], with Razali and Wah [150] finding that the former measure is typically more powerful at detecting departures from normality than the latter for samples of typical IR size (less than 300 topics). These tests are implemented and available using the standard `stats` package in R. Note that the original score distributions are not normal by design as they are bounded between 0 and 1. However, if the risk-adjusted scores are found to be similar in normality properties to the original scores, then the risk-adjusted scores might be considered to be "approximately normal" too. It is important to note that risk-adjusted scores are not required to be normal for the t-test to be valid, as the t-test is valid if the distribution of mean risk-adjusted scores is "approximately normal".

As well as the above normality tests, statistical practitioners assessing the normality of a variable using the SPSS suite of tools [38] also report third and fourth moments of the normal: *skewness* and *kurtosis*; as such, these measures are also employed in this work. Skewness measures how asymmetrical a distribution is, whereas kurtosis measures the how heavily or lightly tailed a distribution is (compared to Gaussian) due to the presence of outliers. The sample skewness [106] of a set X is computed as:

$$\mathit{Skewness}(X) = \frac{\frac{1}{|X|}\sum_{x\in X}(x-\overline{X})^3}{\sqrt{(\frac{1}{|X|}\sum_{x\in X}(x-\overline{X})^2)^3}}\,, \tag{3.14}$$

and kurtosis is computed using Pearson's definition:

$$\mathit{Kurtosis}(X) = |X|\frac{\sum_{x\in X}(x-\overline{X})^4}{(\sum_{x\in X}(x-\overline{X})^2)^2}\,. \tag{3.15}$$

As the kurtosis of the normal distribution has a value of 3, this chapter reports *excess kurtosis* as is typically done in many statistical analyses concerning normality [111], where 3 is subtracted from the kurtosis computed in Equation (3.15). As a rule of thumb, George and Mallery [91] note that values between -2 and 2 for skewness and excess kurtosis are considered to be within the ranges of a standard normal distribution. Values outside of these regions should not be interpreted in isolation as strong evidence for deviation from normality, but should instead be regarded as a warning sign that further investigation is required.

Algorithm 3.1 on the previous page describes the procedure used to observe the relationship between the risk-level r applied with the standard linear risk value function against normality, over many pairs of systems. Where a matrix of per-topic scored runs including their submitting group is formed using the 2012 TREC Web Track collection S_{ijk} using the official ERR@20 metric, $P = 50$ random system pairs are compared for each risk-level of $\mathcal{R} = [1, 2, \ldots, 10]$. Note that the lowest scoring 25% of runs by ERR@20 were filtered out to avoid poorly performing runs affecting the analysis. Similarly, each random system pair is

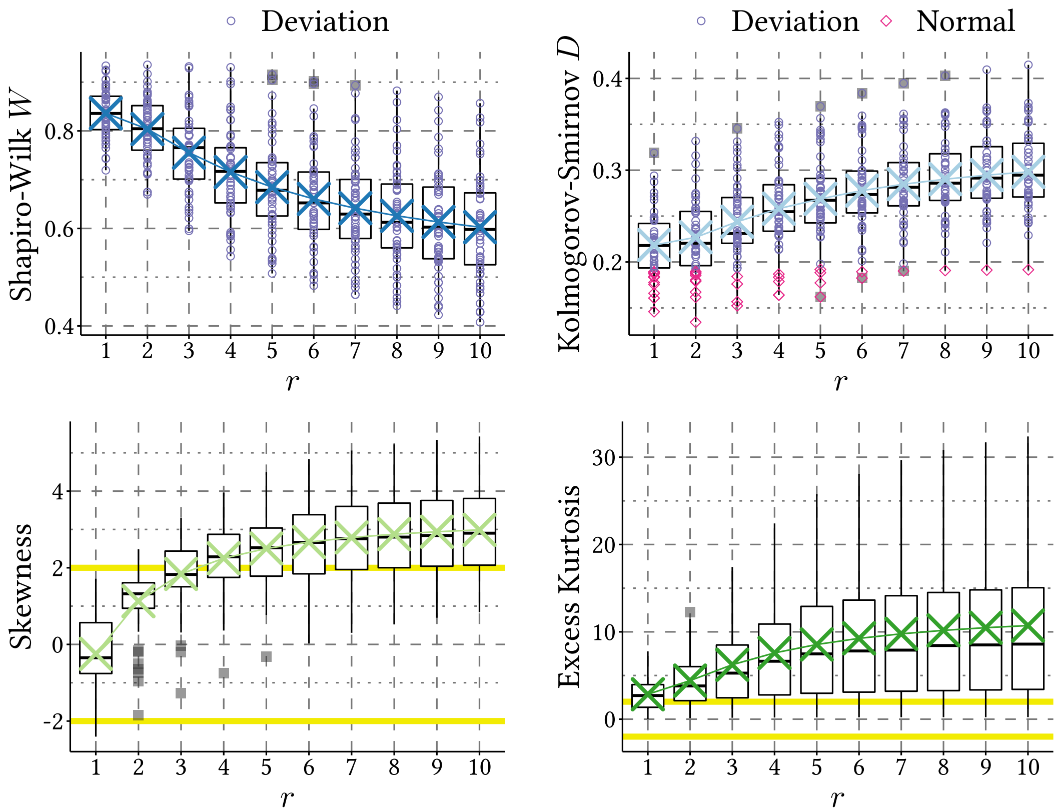


Figure 3.7: Distributional properties of 50 randomly selected paired risk-adjusted scores per risk-level using ERR@20 for 2012 TREC Web Track runs, where each system comparison is a run from a different group (leave-one-out). As r increases the evidence supporting normality decreases. Where a test has a p-value above 0.05, it is marked as “Normal” with a diamond shape, and where deviation from normality is suspected it is marked as “Deviation” with a circle. The region delineated by yellow lines correspond to the rules of thumb of George and Mallery [91] for showing a distribution is normal.

selected with the constraint that only runs from different groups can be compared to avoid intra-group runs skewing results. After the above constraints are met, the score differences are computed and stored in *diff* before applying the risk transformations. The loop beginning on line 8 of Algorithm 3.1 iterates through each of the desired risk-levels in $\mathcal{R}$, applying the transformation on the score differences in a temporary array, which is then instrumented by the above series of normality measures and stored in a table for post-processing. As 49 different systems over a range of r values are now considered, it is no longer practical to present Q-Q plots at this scale to assess the normality of the mean risk-adjusted scores, where a more expensive quantitative approach is required such as those of Parapar et al. [145] or Urbano et al. [197].

The results of that experiment are presented in Figure 3.7. The top-two quadrants correspond to normality tests with an associated p-value, and the bottom two-quadrants are moments of the standard normal distribution for each of the risk-adjusted pairwise comparisons for varying levels of r. The mean value for each normality measurement of the r adjusted values is represented with a cross. As values from the normality tests in the top row have an associated p-value for determining whether the deviation from normality is significant, points indicating a departure from normality are represented with a circle, and conversely, normality is shown using a diamond shape.

The power of the normality tests in the top row reflect the results in Razali and Wah [150], where the Shapiro-Wilk test detects more deviations from normality than the Kolmogorov-Smirnov test. Surprisingly, every Shapiro-Wilk test outcome indicates a departure from normality; even in the cases where no risk adjustment was applied. However, as r increases, the W statistic indicates further deviation from normality. The results of the Kolmogorov-Smirnov test are more in line with expectations, where when no risk adjustment is applied, many test outcomes favor the decision of normality. As r increases, the number of significant detections of non-normality grows proportionately. In the bottom row, moments of the risk-adjusted sample of scores are instrumented under the assumption of a standard normal. The mean skewness enlarges logarithmically and exceeds typical variation when $r = 4$. When evaluating the mean excess kurtosis in the context of the acceptable range of -2 to 2, even before the risk adjustment is applied the excess kurtosis is outside of the acceptable range. That may explain why tests in the top-two quadrants rejected the assumption of normality. However, excess kurtosis worsens logarithmically with respect to r.

In summary, the Shapiro-Wilk W test was too sensitive in its interpretation of deviation from normality, surprisingly declaring any typical ERR@20 score non-normal. However, the results of the Kolmogorov-Smirnov test showed that as r increases, there are more risk-adjusted random pairs of system comparisons that are decidedly non-normal. The raw statistics on every measure show a more complete picture of the distributional properties: as r increases, normality decreases. The magnitude of r may have consequences on the type of statistical test that should be employed on risk-adjusted scores, which is explored next in Section 3.3.

3.2.3 Summary

When exploring the properties of risk-adjusted scores output with the linear and smooth methods against no risk, the $s(\Delta)$ score distribution has a moderate right-tail, but the standard (no risk) and $l(\Delta)$ functions generated points along the reference lines when $r = 2$; partially answering S-RQ3.2. Although the linear approach appeared to be amenable to a normal distribution in Figure 3.6 on page 75, many studies involving risk-adjusted scores employ r values above 2, and the outcome of $s(\Delta)$ showed that the adjustment produced distributions of scores with a concerning deviation from normality. As the smooth function did not demonstrate an advantage over the linear function, the linear function is used hereafter. The finding, however, that the smooth value function impacted the distributional properties of the risk-adjusted scores motivated a more extensive investigation into the original risk function. That experiment involved 50 pairs of randomly sampled systems over many normality measures for varying levels of r, which found that as r increases, the reliability of assuming normality decreases.

Nevertheless, the investigation only considered one collection and metric, and the Student t-test is presumed to be robust to approximately normal distributions. The interpretation of the analysis on risk-adjusted scores is that there is a cause for concern where the deviation from the normality of traditional risk-adjusted scores may be enough to skew statistical inferences. The next section explores how varying r affects inferences derived from tests with different assumptions and makes recommendations concerning paired testing scenarios.

3.3 Risk Inference

Section 3.2 highlighted potential deviations from normality caused by exclusively scaling the losses part of a score distribution, warning that this property may cause parametric inferences to be skewed, to answer RQ3: *How do the distributional properties of risk-adjusted scores impact the results of parametric statistical tests?* Here that assertion is challenged through interpreting the stability of confidence intervals on risk-adjusted scores formed using various techniques via bootstrapping, to answer S-RQ3.3: *How do parametric and nonparametric confidence intervals differ when performing risk-based evaluation?* To show that the effects observed in Section 3.2 are not limited to the selection of ERR@20 as a metric to the TREC web track collections, the official AP metric on two newswire collections, Robust04 and the TREC CORE 2017 *New York Times* corpus is now considered. Additionally, how confidence interval consensus changes due to the number of topics evaluated is also explored, as larger sample sizes influence the activation of the central limit theorem. If normality increases, the confidence intervals should align with each other more closely, addressing S-RQ3.4: *How does evaluating larger topic samples affect confidence intervals when performing risk-based evaluation?*

3.3.1 Selecting A Test

In preparation for the experiments ahead involving different kinds of confidence intervals used for inferential purposes, the reader is reminded of the many kinds of statistical tests available to suit different kinds of data. The key premise is that IR score differences are generally accepted to be approximately normal enough for parametric testing, but the weighted risk-adjusted score differences might not be, with that risk increasing as r increases. Nayak and Hazra [142] describe the necessary steps for choosing the right statistical test based on the data, shown in Figure 3.8 on the following page. While the overall goal of this thesis is to contribute a principled mechanism to perform inferential risk evaluations across many systems at a time, the case of two systems is initially considered, building from first principles towards the multiple-system inference case. On the decision tree, IR effectiveness scores are considered to be numerical. If the experimental results on the traditional r values applied indicate that practitioners should branch left at “Parametric”, then the continued use of TRisk^{-}

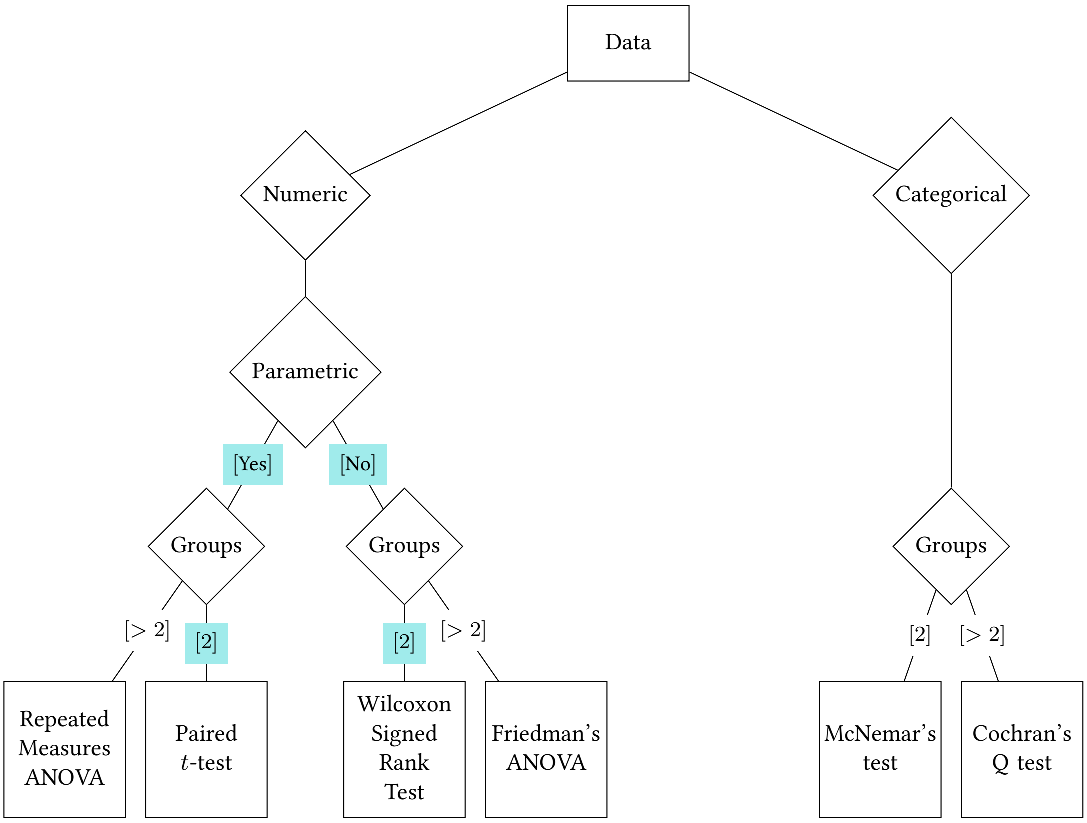


Figure 3.8: An adaptation of Nayak and Hazra [142, Figure 2] which shows a decision tree for the recommended kind of test to use for paired data with various properties. IR effectiveness scores over system-topic scores are numerical, where the question explored in this work is whether the "Parametric" diamond should branch left or right for risk-adjusted scores, for paired data between "[2]" groups.

is justified and its inferences can be safely observed. However, if evidence against normality exists, then Nayak and Hazra [142] recommend using the nonparametric Wilcoxon signed-rank test. Urbano et al. [198] notes that this notion of being nonparametric in the context of a difference in means is not strictly true, as the Wilcoxon signed-rank test assumes that the distribution of the differences is symmetric about the median. Therefore the Wilcoxon signed-rank test is not completely free of distributional assumptions, and this is particularly dangerous when exploring risk-adjusted scores due to the asymmetric adjustments made.

Beyond the Nayak and Hazra [142] recommendations for tests that generate p-values conditioned on effect size and sample size (where 0.05 is the typical threshold for rejecting the null hypothesis) Sakai [158] advocates reporting confidence intervals and effect sizes where possible, instead of just p-values. The bootstrapping the values has two key benefits for the risk-adjusted scores in question, as the bootstrap is completely free of distributional assumptions. Firstly, the shape of the bootstrap distribution can be used as a surrogate for the sampling distribution in a complementary way to the previously explored Q-Q plot mechanism in Figure 3.6 on page 75. And secondly, confidence intervals from the bootstrap distribution for null hypothesis statistical testing [156] can be generated with parametric and nonparametric assumptions. That combined reasoning motivates the use of bootstrap confidence intervals for null hypothesis testing in this work. While the bootstrap statistical test may be more powerful than the confidence interval approach depending on whether the test is one or two-tailed, the confidence interval approach provides effect size and shape information. Additionally, despite randomization tests also being distribution-free, only p-values are generated by them, which hinders their ability to report effect size and shape information. These resampling tests might detect statistically significant differences more readily than the bootstrap confidence interval approach, however, that is left for future work to explore.

Another factor that may play a role in the validity of a test is the significance level used. For example, Colquhoun [57] notes that researchers use evidence to support the claim that an effect exists where they accept that they may be wrong 5% of the time (95% confidence). However, if there is only a practical effect present 10% of the time, the researcher is wrong 36% of the time due to the false discovery rate. Colquhoun [57, p. 12] suggests that to avoid making incorrect claims, "*do not regard anything greater than $p < 0.001$ as a demonstration that you have discovered something*", suggesting the use of a 99.9% confidence interval. As the tails of risk-adjusted scores may deviate from normality, extending the width of the statistical region might exacerbate issues with the inferences derived.

3.3.2 Methodology

The five significance testing approaches used in this chapter to explore S-RQ3.3 and S-RQ3.4 are now defined below, each with differing beliefs on how the population is distributed. The first uses the Student t-distribution to correspond to the TRisk^{-} inferences, and then another four are found using the bootstrap distribution of a statistic [62], using the `R` package `boot` to compute each of the intervals. Note that the goal of this experiment is not to show that any particular combination of systems are indeed significantly different, but rather to observe how parametric and nonparametric testing changes the uncertainty in confidence intervals for varying levels of r, seeking to answer S-RQ3.3.

Student-t Distribution CI. Student-t CIs can be computed by:

$$\hat{\theta}_{\alpha} = t - Q(\alpha, n-1)\frac{s_t}{\sqrt{n}} \text{ and } \hat{\theta}_{1-\alpha} = t + Q(\alpha, n-1)\frac{s_t}{\sqrt{n}}\,, \tag{3.16}$$

where t is the sample statistic, s_t is the standard deviation of the sample, n is the sample size, and $Q(\alpha, \lambda)$ is the quantile function of the Student-t distribution. Quantiles are computed using the `R` function `qt` in the standard stats package. The Student-t distribution corresponds to the confidence limits in TRisk^{-}.

Basic Bootstrap CI. The basic bootstrap method[6] places the confidence interval within the context of the bootstrap distribution and the evaluated sample statistic t (not to be confused with the t-test statistic, which is not used in this particular method). It is basic in the sense that because the distribution is assumed to be symmetrical, it does not require as many bootstrap replicates as other methods.

The distribution of bootstrap replicates t^* reflect the statistic T, whose value in the sample is t:

$$\hat{\theta}_\alpha = 2t - t^*_{((B+1)(1-\alpha))} \text{ and } \hat{\theta}_{1-\alpha} = 2t - t^*_{((B+1)\alpha)} \,. \tag{3.17}$$

The t^* values are sorted in ascending order, and the α and $1-\alpha$ percentiles are used to form the confidence interval. Note that further down in this section, the percentile method is defined which produces confidence intervals entirely from this t^* distribution, where that method is the most commonly used IR approach.

By doubling the mean of the sample statistic t, the confidence interval is based on the assumption that the bootstrap distribution is centered around the sample statistic. For example, assuming a sample with mean 1 is drawn from a normal distribution with mean 1 and standard deviation 1, the 95% confidence interval is $2\times 1-[-1.01, 2.99] = [-0.99, 3.01]$. Now suppose the sample mean is 0.6, but the bootstrap distribution follows a beta distribution with shape parameters 5 and 2. The 95% confidence interval using the basic bootstrap method is $2\times 0.6-[0.35, 0.95] = [0.25, 0.85]$, which is not centered around the sample mean 0.6 and yields an asymmetric interval (the sample mean would need to be 0.55 to be centered).

Student-t Distribution Bootstrap CI. Another parametric approach, this takes the basic CI form, except the $N(0, 1)$ approximation is replaced with $Z = (T-\theta)/\sqrt{V}$, where V is the variance statistic, and θ is the true value of T to form an error region. For each of the sets formed in the B bootstrap iterations, the t^* and v^* values are used to compute the set of replicates that have been standardized into a z-score $z^* = (t^*-t)/\sqrt{v^*}$:

$$\hat{\theta}_\alpha = t - z^*_{((B+1)\alpha)}\sqrt{v} \text{ and } \hat{\theta}_{1-\alpha} = t - z^*_{((B+1)(1-\alpha))}\sqrt{v} \,, \tag{3.18}$$

where v is the variance of the original sample.

Percentile Bootstrap CI. The percentile method can be used on parametric and nonparametric bootstrap samples. Supposing bootstrap replicates t^* are sorted:

$$\hat{\theta}_\alpha = t^*_{((B+1)\alpha)} \text{ and } \hat{\theta}_{1-\alpha} = t^*_{((B+1)(1-\alpha))} \,, \tag{3.19}$$

the α and $1-\alpha$ percentiles are used to form the confidence interval.

[6] A reviewer has noted there could be confusion as to what constitutes the *basic* bootstrap, this work uses the Davison and Hinkley [62] interpretation.

Algorithm 3.2: The BCa^{-} method of performing risk-inference on IR scores.

```
Input: An array of IR effectiveness scores for an experimental system $s_1$ and
       baseline $s_2$, with a required risk-level $r$, for $B$ bootstrap replicates, where $T$
       is the statistic on the paired risk-adjusted scores (typically the mean), and $\alpha$
       is the required significance level.
Output: The confidence intervals of the bias-corrected accelerated Bootstrap
        estimates of $T$ on the risk-adjusted score differences.
$\Delta \leftarrow s_1 - s_2$
$\Delta \leftarrow -1 \times \Delta$
$\Delta_r \leftarrow \{\Delta^{+} \cdot r\} \cup \{\Delta^{-}\}$                          // Perform risk-adjustment
$t \leftarrow T(\Delta_r)$                                                  // Statistic on original sample
                                                      // Generate $B$ Bootstrap replicates for $T$
$t^{*} \leftarrow$ array$(B)$
$w \leftarrow 0$                                                          // Bias correction parameter
for $b \in B$ do
    $c \leftarrow$ random_sample$(\Delta_r$, replacement=true$)$
    $t^{*}[b] \leftarrow T(c)$                                           // Statistic on bootstrap sample
    if $t^{*}[b] < t$ then
        $w \leftarrow w + 1$
    end
end
$w \leftarrow w/B$                                                        // Bias correction parameter
$z_{\alpha} \leftarrow \Phi^{-1}(\alpha)$                          // Transform sig. level to $z$-score of std. normal
                                          // Jackknife $T$ using $\Delta_r$ to compute acceleration factor
$l^{*} \leftarrow$ array$(|\Delta_r|)$
for $n \in |\Delta_r|$ do
    $\Delta_r' \leftarrow$ subset$(\Delta_r, 0, n)$                                            // Subset $\Delta_r$
    $l^{*}[n] \leftarrow T(\Delta_r')$                              // Compute Jackknife $T$ estimate on $\Delta_r'$
end
$L \leftarrow (|\Delta_r| - 1)(\overline{l^{*}} - l^{*})$                          // Bias-corrected Jackknife Estimate
$a \leftarrow sum(L^3)/6\left[sum(L^2)\right]^{3/2}$                        // Bias-corrected Acceleration factor
                       // Bias-corrected accelerated significance level supplied to Basic CI
$\tilde{\alpha} \leftarrow \Phi\left(w + (w + z_{\alpha})/(1 - a(w + z_{\alpha}))\right)$
$\hat{\theta}_{\tilde{\alpha}} \leftarrow 2t - t^{*}_{((B+1)\tilde{\alpha})}$                       // Lower CI – Lower Bootstrap Quartile
$\hat{\theta}_{1-\tilde{\alpha}} \leftarrow 2t - t^{*}_{((B+1)(1-\tilde{\alpha}))}$                  // Upper CI – Upper Bootstrap Quartile
return $(\hat{\theta}_{\tilde{\alpha}}, \hat{\theta}_{1-\tilde{\alpha}})$
```

Collection	Citation	Documents	Unique Terms	Total Terms	Topics
Robust04	[199]	528,155	664,603	253,367,449	250
TREC17	[13]	1,855,658	2,970,013	1,285,653,766	50

Table 3.4: Collection statistics of the Robust04 and TREC17 collections used to explore the agreement of parametric and nonparametric statistical testing approaches for risk evaluation.

Bias-Corrected and Accelerated Bootstrap CI (BCa^{-}). If the swap of quantile estimates comes from a biased scale the results will be inaccurate. To address that for the percentile method on nonparametric bootstrap samples, calculating and correcting for the bias in the distribution is needed. Using t^{*}, the bias-correction parameter w can be computed, which corresponds to the ratio of the number of times a t^{*} value is less than t, to B. These ratios are then then transformed into a z-value. The quantile function of the normal distribution is again used to transform the upper and lower α confidence limits to z-values denoted z_{α}. Then, the acceleration parameter a is computed. A set of l^{*} Jackknife[7] estimates of T is formed, where the set L corresponds to the difference in the mean Jackknife estimate and the corresponding value computed in l^{*}. Acceleration a is defined as:

$$a = \frac{\sum_{l \in L} l^{3}}{6 \left(\sum_{l \in L} l^{2}\right)^{3/2}} . \tag{3.20}$$

The adjusted z_{α} values are now used to compute the adjusted CI $\tilde{\alpha}$ using w and a, that will be used in the basic bootstrap CI formula listed in Equation (3.17) on page 83:

$$\tilde{\alpha} = \Phi \left(w + \frac{w + z_{\alpha}}{1 - a(w + z_{\alpha})} \right) , \tag{3.21}$$

where Φ is the cumulative distribution function of the standard normal distribution. After substituting $\tilde{\alpha}$ for α in Equation (3.17) on page 83, the BCa^{-} CI has been computed. The methodology to compute the BCa^{-} method on risk-adjusted IR scores is formalized in Algorithm 3.2 on the previous page.

3.3.3 Experimental Setup

Different corpora and risk baselines to that of Section 3.1.2 on page 65 are now explored to ensure risk-adjusted score properties are generalizable across collections, to fully answer research questions S-RQ3.3 and S-RQ3.4.

Corpora and Risk Baselines. The investigation into risk score distributions continues now from an inferential perspective, using the TREC *New York Times* CORE 2017 corpus (TREC17) [13], as well as the Robust04 corpus (Robust04) [199], where TREC17 has the usual 50 topics and Robust04 has 249 (250 originally, but topic 672 has no relevant documents and

[7] Jackknife resampling Wikipedia article: `https://en.wikipedia.org/wiki/Jackknife_resampling`, accessed on 3rd November 2022.

is hence dropped). Table 3.4 on the previous page describes summary statistics for these comparable newswire document collections. The Robust04 collection is one of the most widely used collections in IR, known for having a substantial number of topics, and relatively deep judgments. These additional topics will be useful to answer S-RQ3.4.

To explore the inferential behavior of risk on varying topic sizes on the Robust04 collection, where the typical count of topics in an IR test collection is 50, the Robust04 set is sampled by taking every fifth topic, for example, 301, 306, ..., 696. (Other samples of 50 were also measured, with similar results to those about to be presented.) Then, the full 249 Robust04 topic set is evaluated, where the difference in support for the various significance tests listed above is interpreted. For each of the Robust04 and TREC17 corpora, the champion system is fixed as an untuned Okapi BM25 run computed using Indri 5.11. That run differs from the `indriCASP` run in Section 3.1.2 on page 65 where query likelihood is used with a spam filter.

Significance Levels. Colquhoun [57] advises statistical practitioners to report a 99.9% confidence interval alongside the commonly reported 95% interval to avoid false-positives. In following that advice, both the 95% and 99.9% intervals are considered when examining how that affects inferences of risk-adjusted scores.

Bootstrapping. As conducting a hypothesis test using the Bootstrap requires there to be a set number of iterations, 100,000 score replicates are used in this study, in line with previous IR experiments [145, 179, 197]. For 249 topics, the number of bootstrap replicates is proportionally increased to 500,000. The same set of score replicates are used to compute each of the four different kinds of bootstrap confidence interval explained in the methodology section above, to allow their inferences to be fairly compared.

3.3.4 Experimental Analysis

Head-To-Head Risk CI Agreement. Recall the Hesterberg et al. [100] observation that the shape of a sample statistic distribution can be examined when plotting the density of the bootstrap replicates. If these risk-adjusted values are not distributed symmetrically, then parametric tests may have weakened accuracy. Figure 3.9 on the following page shows the shapes of each of the risk-adjusted score distributions when varying r, comparing the best-submitted run in the Robust04 track (`pircRB04td2`) against the champion BM25 run from Indri 5.11. The first element in each pair is a density plot of the bootstrapped risk score. Then the confidence intervals for each test are shown below each density plot. The `Student` interval corresponds to the Student t-test when $r = 1$, and when $r > 1$ the inferences derived using TRisk^{-}. When $r = 1$, gains and losses are treated equally, and the CIs are practically equal between tests. As r increases, the CIs increasingly disagree. These changes can be characterized by the general observations for increases in r:

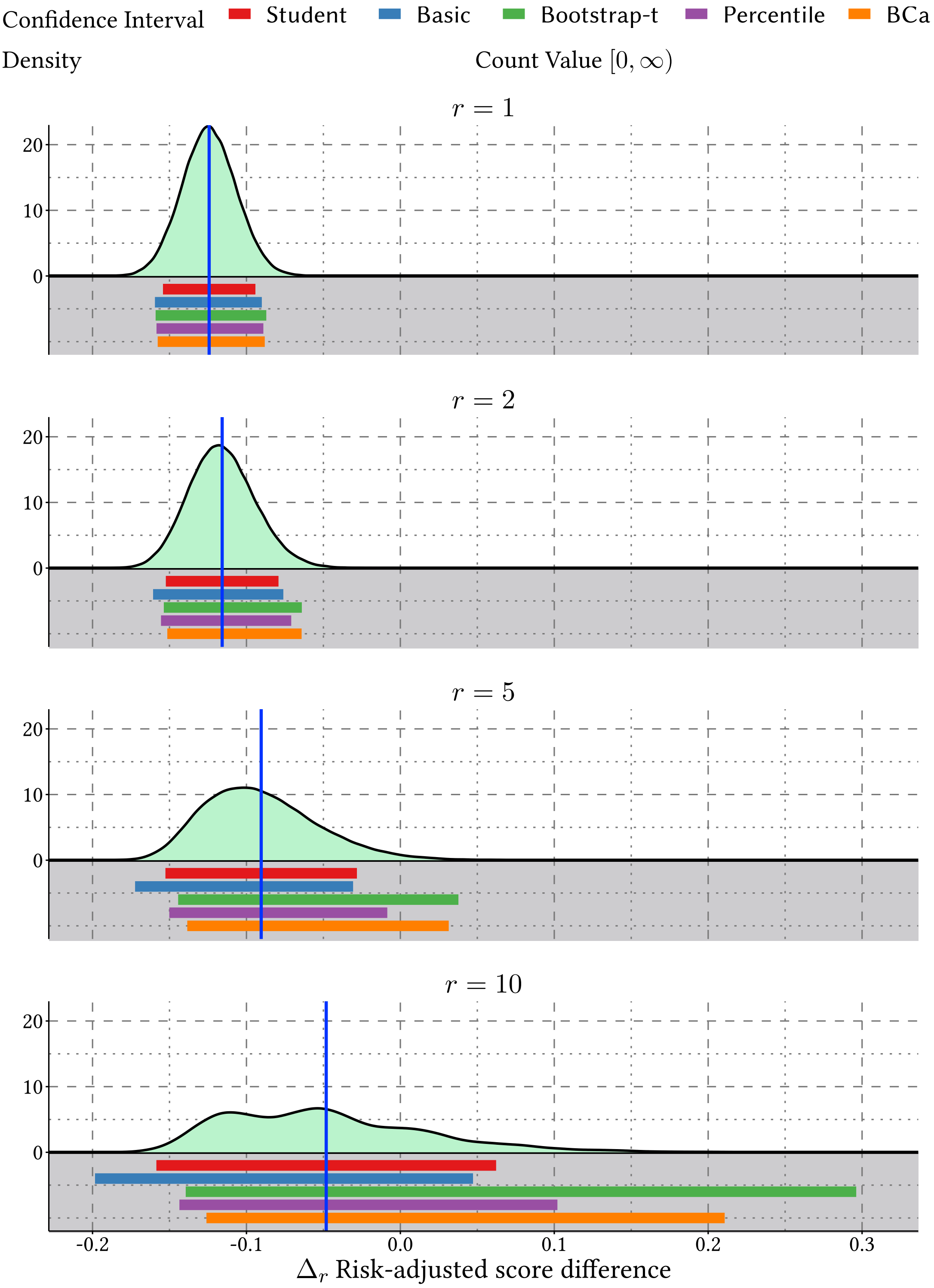


Figure 3.9: The shape of the URisk$^-$ bootstrap score replicates on the AP metric when a BM25 baseline is compared against the best-submitted run to the Robust04 track, `pircRB04td2`, measured by AP. As r increases, the shape of the bootstrap sampling distribution shifts to a skewed shape, which is reflected in the various confidence intervals disagreeing as parametric assumptions erode.

- Student – Increases in risk do not reflect the shape of the bootstrap replicates. When $r = 5$, larger risk values show right-skew in the bootstrap distribution, but the left and right bounds on the sample mean (blue line) are equal in width.

- Basic – This bootstrap CI assumes risk-adjusted scores are Gaussian. As r increases, it appears that this CI will be the most difficult to achieve significant reward (no chance of risk), however, the right fence of the CI tends to follow the Student method closely, which may lend to over-reporting significant risk.

- Bootstrap-t – As one of the more extreme parametric bootstrap CIs considered, it assumes the bootstrap distribution follows a Student t-distribution, which may be a safe assumption when $r \in 1, 2$, but grows disproportionately when $r \in 5, 10$.

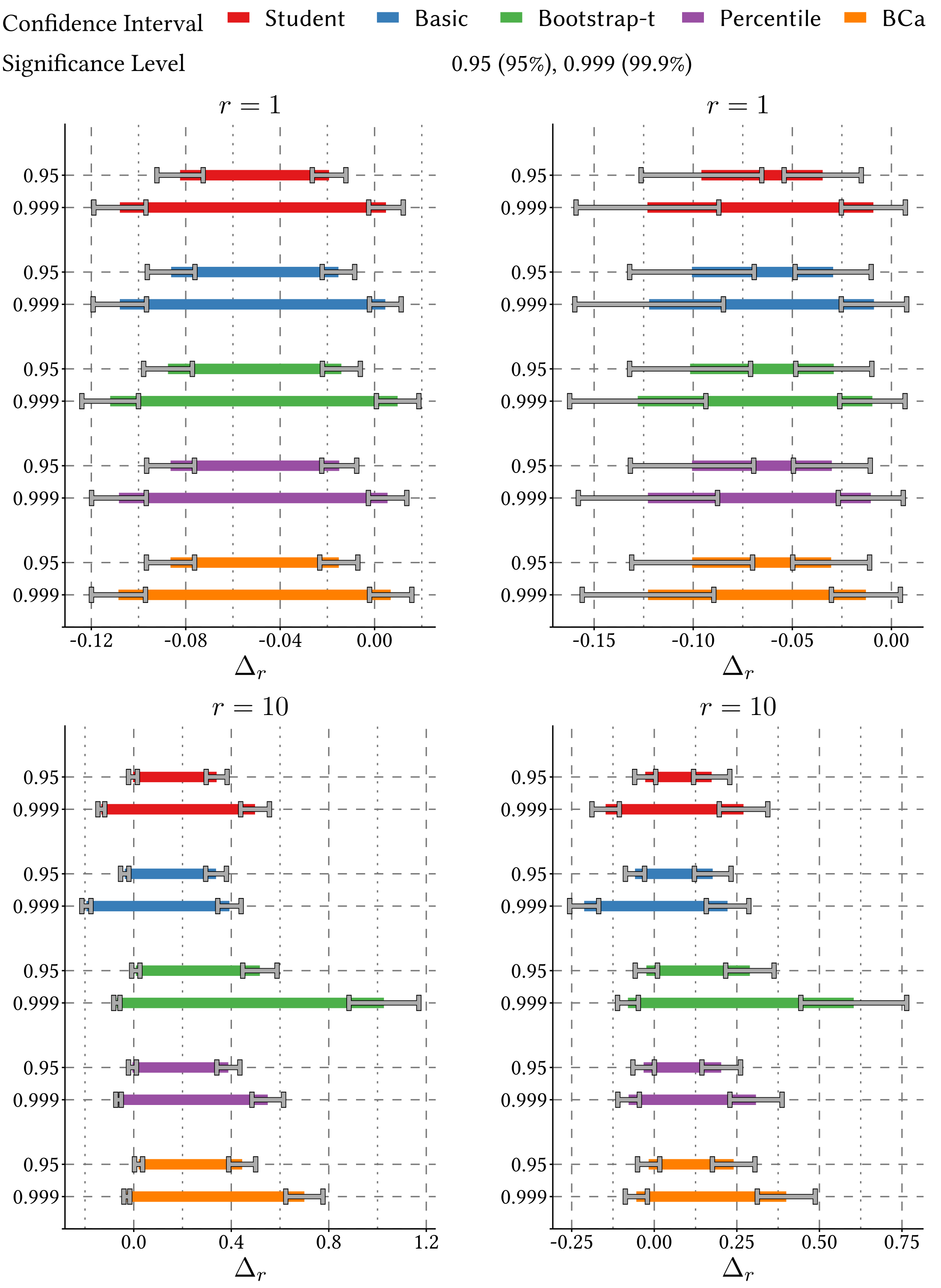
- Percentile – This method is the most commonly employed Bootstrap CI used. It is nonparametric in the sense that no particular distribution is assumed, however, it does assume that the values are distributed with symmetry. The right-shifted interval produced has a similar width to the Student and Basic CIs.

- BCa – This bias-corrected nonparametric CI is the most right-shifted of the set, however, the upper bound is comparable to the Bootstrap-t interval. The BCa^{-} method is the most powerful approach of the methods considered in this experiment when finding no chance of risk, and one of the harder tests to find a significant chance of risk.

The divergence in CI agreement can be attributed to weighting one side of the distribution more than the gains side, which erodes the parametric assumption when larger r values are considered. A more exhaustive analysis on both Robust04 and TREC17 is now conducted to fully answer S-RQ3.3.

Exhaustive Analysis. The spread of CIs generated when many systems are compared is now observed. As the AP measure is now in use, the bottom 25% of runs are discarded by AP to avoid erroneous runs skewing results. The remaining runs submitted to the TREC17 and Robust04 tracks are compared as challengers to the baseline BM25 run. For each paired comparison, the CIs at 95% and 99.9% significance levels are computed, with no risk bias ($r = 1$) and high-risk bias ($r = 10$), with each of the CIs computed. For each collection, significance level, and test triplet, the median lower and upper ends of the system comparisons are presented with a solid colored bar. For each of these median values, a confidence interval is presented indicating the 95% confidence intervals. Those intervals are computed using the McGill method which utilizes the CI interquartile ranges from all systems in the comparison.

Figure 3.10 on the following page plots the results. In both $r = 1$ panes for the 95% confidence intervals, the median upper fence of the CI distributions indicates that, on average, systems outperformed the baseline BM25 run (negative score differences indicate reward); however, when $r = 10$ the risk values no longer indicate that other systems are superior. But, most importantly, when $r = 10$ the parametric and nonparametric CI methods disagree,



(a) Robust04 50 sub-sampled topics.

(b) TREC17 corpus.

Figure 3.10: How CIs change with r and significance level. In each graph the top 75% of submitted runs are compared against a BM25 baseline using AP and r. The median CI limits (over systems) are plotted as the left-hand and right-hand ends of each solid horizontal bar, and 95% intervals showing the CI limits' ranges are added. Note the different horizontal scales on the panes.

which is exacerbated when the larger 99.9% significance level is considered. When $r = 10$, the parametric assumptions are violated when observing the density plot in Figure 3.9 on page 87, where the CIs generated using nonparametric methods better capture this information for larger r values. Note, however, that there is no ground truth for a right or wrong statistical inference. Many practitioners choose to report two significance levels to increase confidence in their results.

Normality Concerning Topic Count. To answer S-RQ3.4, a larger topic set is considered to determine whether one of the five approaches yields more trustworthy parametric inferences. The previous explorations all showed disagreement in the confidence intervals when the topic sample size was 50. However, when a sufficiently large topic set is available, the CIs may converge towards normality due to the central limit theorem. To observe that effect taking place, the equivalent exhaustive experiment on 249 topics is presented in Figure 3.11 on the following page, to be contrasted against the 50-topic plot in Figure 3.10a on the previous page. There is a greater agreement between the nonparametric and parametric tests when larger topic set sizes are considered, but note this is against one baseline system.

Previously, Figure 3.7 on page 78 presented a more expansive analysis using ERR@20 on the 2012 TREC Web Track data to explore normality, where the baseline and experimental pair were randomly selected. A more expansive analysis is now presented with 50 and 249 topics in Figure 3.12 on page 92, using Robust04 and the AP metric, comparing the normality of risk-adjusted scores on 50 and 249 topics. Recall from Algorithm 3.1 on page 76 that the system pairs are randomly selected, where the lowest scoring 25% of runs by ERR@20 were filtered out and only runs from different groups are compared. When both the experiment and baseline systems are free to vary, the Shapiro-Wilk W statistic indicates that on average the normality assumption is less certain when 249 topics are used instead of 50 on the risk-adjusted score distributions. However, it is not how the risk-adjusted scores are distributed that is of great importance, but rather how the mean differences between the risk-adjusted scores are distributed. Figure 3.11 on the following page shows that the differences in confidence interval lengths between different tests are reduced when 249 topics are used instead of 50, suggesting that the central limit theorem is at play for larger sample sizes.

3.3.5 Discussion

This section explored how confidence intervals change based on different risk computations in two different collections, Robust04 and TREC17. The results show that when typical IR collections of 50 topics are used to perform inferential risk-biased evaluations, distribution-free resampling based nonparametric tests should be preferred, addressing S-RQ3.3. Interestingly, the disagreement between nonparametric and parametric confidence intervals reduced using the Robust04 collection using 249 topics when the BM25 baseline was fixed, despite how the risk-adjusted scores were distributed. Erring on the side of caution, the experiments do appear

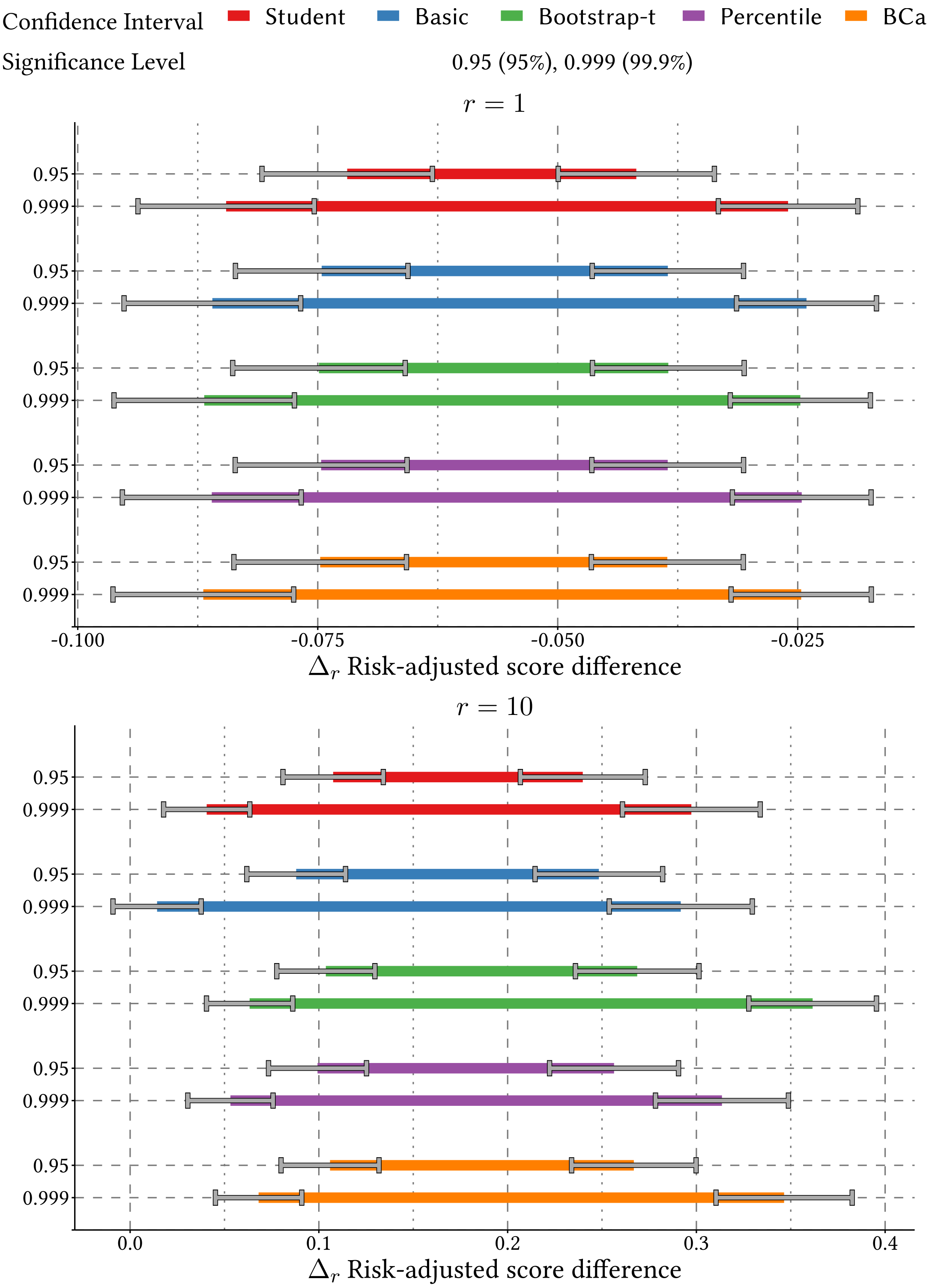


Figure 3.11: A plot of the spread of CIs generated using the larger 249 topic set of Robust04, to be compared against the 50 topic set in Figure 3.10a on page 89. When compared with Figure 3.10a, more topics tend to increase the agreement among nonparametric and parametric CIs for larger values of r due to the effects of the central limit theorem.

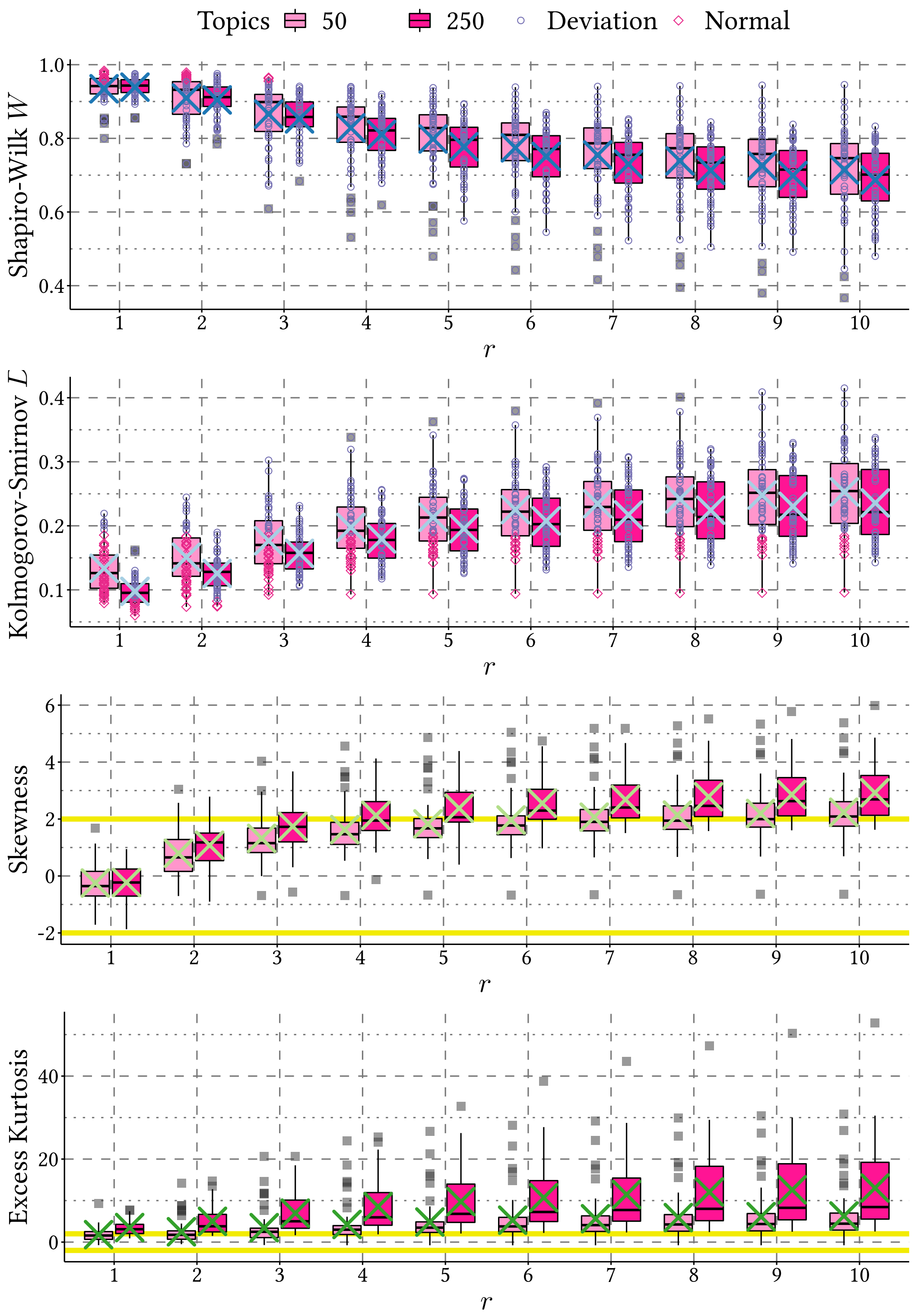


Figure 3.12: Distributional properties of 50 randomly selected paired risk-adjusted scores per risk-level using AP and the Robust04 runs, where each system comparison is a run from a different group (leave-one-out). The 50 topics are a subset of the full 249 topic-set, where the same pairs of systems are compared. As r increases the evidence supporting normality decreases. Tests with a p-value above 0.05 are denoted “Normal” with a diamond shape; otherwise it is marked as “Deviation” with a circle.

to support using nonparametric tests over parametric ones on risk-adjusted scores regardless of whether 249 topics are used instead of the more typical 50 topic collection size; answering S-RQ3.4.

Although the generality of nonparametric tests has the cost of reduced statistical power, the accuracy of parametric tests may require justification on risk-adjusted scores in light of these findings, which is of particular concern when large r values are used. Of the range of nonparametric tests available, the results in this section provide a good rationale for using the bias-corrected accelerated bootstrap (BCa^{-}) confidence intervals generated for risk inference, as they provide a confidence interval about the specific statistic of interest which provides more information than p-values alone. Other valid nonparametric tests may apply on risk-adjusted IR scores, such as resampling techniques like the Bootstrap test [156] or permutation tests, both of which have been a mainstay in IR in statistical investigations involving standard effectiveness scores [145, 197].

3.4 Conclusion

Whether parametric statistical tests can be faithfully used on paired IR effectiveness score differences has been debated frequently with mixed results. In the absence of a definitive answer to the correctness of a statistical outcome, combined with many positive conclusions drawn, the Student t-test remains the most commonly employed choice for inferential purposes. Motivated by the ubiquity of the t-test, this chapter focuses on the novel angle of exploring the appropriateness of parametric testing on risk-adjusted scores. Given the wealth of positive results about the Student t-test in the IR community, any situation where the test produces irregular results is interesting because it is usually robust to scrutiny.

This chapter began with the goal of building on the existing inferential TRisk measure through exploring the potential for inferential testing with an alternative smooth searcher-inspired risk value function $s(\Delta)$, instead of the standard linear version $l(\Delta)$. To validate that idea, the relative system rankings and distributional properties of these different risk-adjusted score outputs were compared on recent risk overlays in circulation. Surprisingly, the experiments using the smooth searcher-inspired value function $s(\Delta)$ found no supporting evidence that practically different system orderings would be produced compared to the standard $l(\Delta)$ on popular risk overlays, answering S-RQ3.1. When inspecting the distributional properties of the bootstrapped mean values generated between $s(\Delta)$ and $l(\Delta)$ when $r = 2$, the smooth value function produced risk-adjusted scores deviating from a Student distribution. As $r = 2$ is often the lowest r value considered, the usual $l(\Delta)$ function hinted at potential deviations as well for larger r values. A more expansive experiment randomly selecting baselines and experimental systems found that as r increases, the chance of non-normality increases, potentially impacting the outcomes of significance tests which assume normality (or approximate normality), partially answering S-RQ3.2.

Having demonstrated that risk overlays might be subject to more uncertainty around whether the Student t-test can be safely applied with them, the agreement of different confidence intervals of the mean with parametric and nonparametric attributes are explored, investigating the impact of the r risk parameter. As hypothesized, these opposing methods of computing confidence intervals disagreed when the shape of the risk-adjusted distributions became more skewed and deviated from the expected symmetrical bell-curve form, answering S-RQ3.3. Since 250 topics may be more amenable to parametric testing by virtue of the central limit theorem, this was also explored. When the baseline and experimental systems were free to randomly vary, the resulting risk-adjusted distributions deviated from normality more decidedly with a larger topic set, however, the inferences in the mean statistic was more stable due to the central limit theorem, answering S-RQ3.4. These findings suggest that using nonparametric testing methods on risk-adjusted scores may be important on standard topic set sizes if statistical inference is desired, especially when large r values are considered, answering the overarching thesis research question RQ3. In the next chapter, these insights about how risk-adjusted scores can impact inferences are used to investigate how risk could be compared over many systems. The approach explores a fit-for-purpose statistical design that corrects for multiple comparisons.

4

Modeling Risk-Adjusted Scores On Many Systems

Capturing the variability of per-topic gains and losses between challenger and champion systems is important, to ensure that net effectiveness improvement does not get compromised by degrading topics that were effective on the champion system. Chapter 3 on page 59 aimed to better understand the statistical properties of risk-adjusted score differences in a head-to-head testing setting. The key finding from the previous chapter was that the asymmetric risk transformation on the losses part of the distribution may have consequences for inferential techniques. Although standard paired IR effectiveness scores tend to produce consistent inferences with 50 topics [180], the results of the previous chapter suggest that risk-adjusted scores might require larger sample sizes than what is available in public IR datasets before their difference in mean score distributions can be assumed to be "approximately normal". A proposed solution in the paired risk-testing scenario was to use statistical testing approaches that account for skewness when generating confidence intervals used for inference.

As the deviation of risk-adjusted scores from normality in the paired risk scenario poses challenges from an inferential perspective, this chapter explores a novel approach for risk inference involving multiple systems which honors the asymmetry of risk-adjusted score distributions. A core issue whenever multiple system testing is concerned is the problem of correcting for family-wise error. The most common practice towards correcting for multiple comparisons on IR scores is to use ANOVA combined with Tukey's HSD. Although the normality assumptions of ANOVA is not perceived to be an issue on standard IR scores, the results of the experiments in the previous chapter suggest that deviations from normality may impact inferential results on risk-adjusted scores (especially for larger values of r). An alternative versatile inferential methodology widely accepted in the statistical community is Bayesian inference, made possible by recent advances in software and hardware to run the MCMC simulations Ryan et al. [154]. Bayesian statistics employs researcher-defined prior distributions in combination with likelihoods from the data to form hypotheses. The paradigm has been used in the past [40, 42, 161] to explore paired system evaluations, but as yet no work has looked into the special case of modeling risk over many systems; further, there has not yet been work on standard IR scores with multiple comparison correction in the Bayesian way. Exploring the utility of modeling standard IR scores in a multiple comparison framework is important, as the classic inferential tests tend not to scale with many comparisons [160].

This chapter expands the Chapter 3 findings from paired inferential risk comparisons into multiple system risk comparisons. A novel Bayesian method for inferring over multiple IR systems that is sensitive to differences in effectiveness is explored, where empirical analysis is provided for performing risk inference on pairs of systems, and on many with correction. A key rationale for investigating this is because, rather than investing all engineering resources to produce a single challenger to a champion system, experimenters often wish to evaluate many systems simultaneously. No prior work has explored the statistical properties of risk-adjusted scores in the multiple system setting with multiple comparison correction.

A novel exploration for computing Bayesian inferences on paired risk-adjusted score differences is explored, which provides further evidence to the Chapter 3 asymmetry findings. When modeling the scores using a skew-normal distribution (discussed in Section 2.4), the skewness parameter λ grows as the risk-level increases, indicating that it is a useful parameter for modeling risk-adjusted scores. In showing that risk-adjusted scores can be better described using a Skew-Normal distribution, this motivates exploring how to compare multiple systems in a fully Bayesian approach. *Bayesian hierarchical modeling* (BHM) is conceptualized, a technique yet to be explored in an IR context, that can be used to compare multiple groups of data simultaneously for inferential purposes. After demonstrating some of the operating characteristics of BHM with a small example, it is expanded upon to explore its ability to statistically differentiate a set of runs while exploiting previously shared 'artifact' runs to reduce noise in the system effect estimates used for inference. The new BHM approach is further explored when a risk-adjustment is applied to many of the systems compared to a single system, and the implications of doing that and per-topic risk forecasting is considered.

Contributions. This chapter addresses the thesis research question and sub-questions:

Research Question (RQ4): *How can risk-adjusted scores be modeled over multiple systems with multiple comparison correction?*

S-RQ4.1: *How do Bayesian inferences of risk involving pairs of systems compare when incorporating the asymmetry into the model?*

S-RQ4.2: *How does using previous system artifacts affect Bayesian inferential results for IR test collections?*

S-RQ4.3: *How does the Bayesian prior affect inferential results using risk-adjusted scores for one-to-many comparisons?*

S-RQ4.4: *How do Bayesian and frequentist credible and confidence intervals differ when performing risk-adjusted evaluations?*

This chapter introduces the first Bayesian approach to computing risk-adjusted inferences on many systems with correction for multiple comparisons.[1] It is also the first Bayesian methodology used in an IR setting for evaluating inferences over many systems with standard IR effectiveness scores. The groundwork is laid for exploring whether a Bayesian approach may be useful for risk-adjusted inferences, integrating asymmetry into the modeling assumptions of paired risk-adjusted scores, as was found to be important in Chapter 3 (S-RQ4.1). Experimental analysis using the Bayesian hierarchical modeling methodology which enables multiple system comparisons, found that the number of artifact systems supplied as reference information to a Bayesian model is a factor in reliability of the generated models, and that the more runs available, the more consistent the inferences are (S-RQ4.2). When multiple systems with a risk-adjustment applied are modeled with a skew-normal distribution, the model is better able to describe the data (S-RQ4.3), however, the ability to achieve statistical significance is weakened compared to the traditional approaches, especially for larger risk r levels (S-RQ4.4). (Chapter 5 follows up on this unexpected reduction in statistical power by exploring leaving the IR scores unadjusted prior to modeling, and computing risk as a posterior predictive statistic on the simulated BHM model.)

Figure 4.1 on the next page signposts the flow of the evolving ideas throughout the chapter. A novel Bayesian approach for computing credible intervals on paired risk-adjusted scores is introduced in Section 4.1, which is used to build towards the methodology for computing risk inference on multiple systems. Section 4.2 commencing on page 104 introduces the concept of Bayesian hierarchical modeling for the first time in an IR evaluation setting, where the framework allows for making inferences over many groups of data with multiple comparison correction. A small-scale example involving a pool of challenger systems and a single champion system is defined, and then expanded upon using several newswire corpora and a pool of previous artifact systems. Section 4.3 starting on page 117 explores simulating a BHM on risk-adjusted scores over many systems with artifact pools, placing the method in context with existing risk overlays (S-RQ4.4). The chapter is concluded in Section 4.4 on page 123, answering the research question and sub-questions.

4.1 Paired Bayesian Risk

Chapter 3 showed that the skewness of risk-adjusted score distributions can violate the symmetric assumptions of statistical tests, and subsequently change the outcomes of statistical tests when 50 topics are considered. This section addresses S-RQ4.1: *How do Bayesian inferences of risk involving pairs of systems compare when incorporating the asymmetry into the model?* As the focus of this chapter is to understand the benefit of applying Bayesian inference as *a* solution to the problem of comparing multiple systems on risk, the paired-testing case is considered first, which is later used to construct a solution for multiple testing in Section 4.3.

[1] This work appeared in R. Benham, B. Carterette, A. Moffat, and J. S. Culpepper. Bayesian Inferential Risk Evaluation on Multiple IR Systems. *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 339–348, 2020.

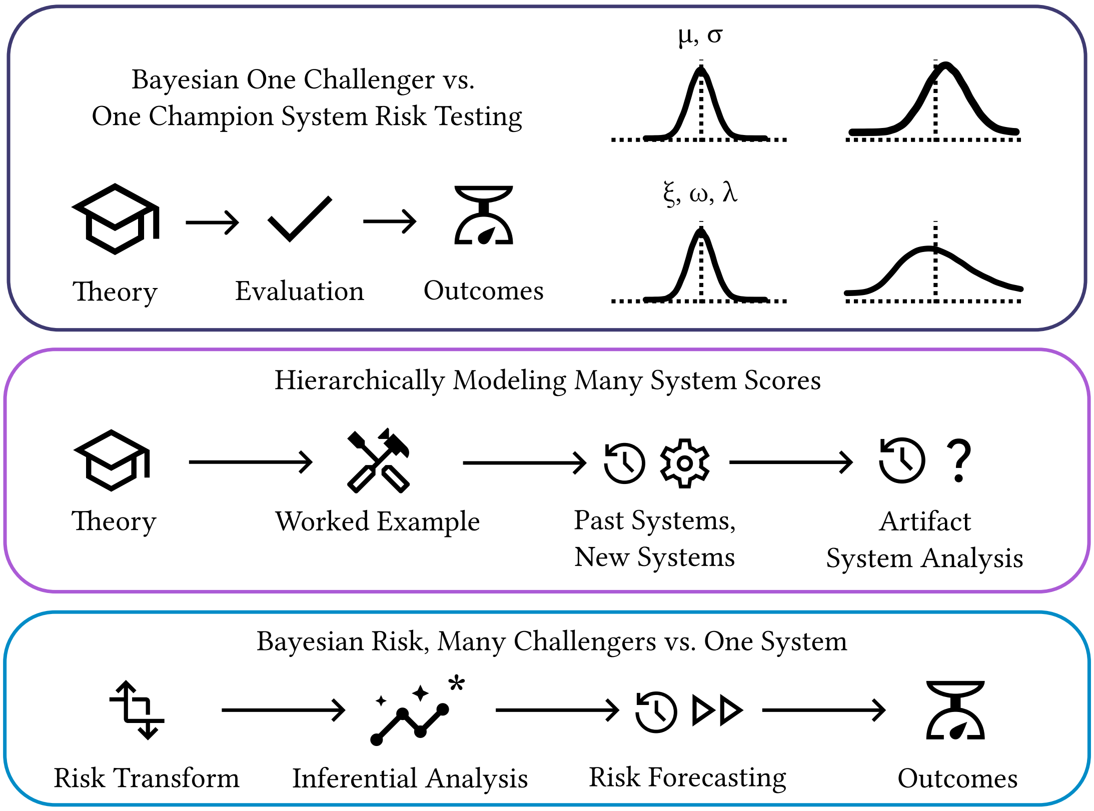


Figure 4.1: A diagram presenting a high-level road-map of the chapter. A Bayesian approach risk inference is explored over pairs of systems, then the ability to perform Bayesian inference over many systems of unadjusted IR scores is explored. Finally, the outcomes of both investigations are merged to explore modeling risk-adjusted scores over many systems for inferential purposes.

The models used in this section are first described, and an experiment is then conducted to compare the inferential outcomes of the Paired-Gaussian and Paired-Skew-Normal approaches on real IR effectiveness data.

4.1.1 Theory

As the results of Chapter 3 showed, skewness in the mean difference in risk-adjusted score distributions can violate the pre-requisite assumptions of statistical tests. Using that insight, modeling the risk-adjusted scores as Skew-Normal is a safe choice, as it is a generalization of the Gaussian distribution that allows for skewness by adding a third shape parameter λ. If in the worst-case the risk-adjusted scores are symmetric and the assumption of skewness is flawed, then the Skew-Normal distribution location and scale is exactly equivalent to a Gaussian distribution. The relationship between the Skew-Normal and the standard Gaussian distribution is described in detail in Section 2.4 on page 48. The Bayesian models for describing IR effectiveness score differences are now described in detail for both the Paired-Gaussian and Paired-Skew-Normal approaches for pairs of systems.

Algorithm 4.1: The Paired-Gaussian posterior log-density accumulator.

```
  Input: The array $\Delta$ of IR effectiveness score differences of two systems, with current
         chain (or, initial) proposals for $\langle b, \sigma \rangle$ used to explore the posterior
         distribution.
  Output: The accumulated log-posterior density for the given inputs, for the MCMC
          sampler to probabilistically determine whether remaining in the current
          position or moving to this location is optimal.
                                      // Accumulate prior density using brms defaults
1 $location \leftarrow median(\Delta)$
2 $scale \leftarrow \max\{mad(\Delta), 2.5\}$
3 $lp \leftarrow student\_t\_lpdf(b, 3, location, scale)$
4 $lp \leftarrow lp + student\_t\_lpdf(\sigma, 3, 0, scale)$
                                      // Accumulate likelihood density
5 for $y \in \Delta$ do
6     $lp \leftarrow lp + normal\_lpdf(y, b, \sigma)$
7 end
8 return $lp$
```

Paired-Gaussian. Algorithm 4.1 shows the steps involved for computing the log-density accumulator supplied to an MCMC sampler for the Paired-Gaussian model on paired IR score differences. This method is used as a baseline for comparison to the Paired-Skew-Normal model on risk-adjusted score differences of pairs of systems, and also forms the basis for the multiple system comparison approach explored in Section 4.2. As the goals of the Paired-Gaussian model are to estimate the mean and variance of the score differences without the need to model the correlation between the two systems, the model is a simplified variant of the multivariate Gaussian approach in Algorithm 2.4 on page 47 used by Sakai [161].

Another key difference between the Sakai [161] model and the Paired-Gaussian model is the definition of a weakly-informative prior on line 3 of Algorithm 4.1. All weakly-informative priors defined in this chapter, and subsequently, this thesis, use the default settings of the R package `brms`; their function is to aid in the convergence of the MCMC sampler. Bürkner [36] notes that the default priors for mean and scale parameters use a half Student-t distribution with 3 degrees of freedom, as it provides better convergence than the half-Cauchy prior. Carterette used weakly-informative priors in their work on Bayesian modeling of IR effectiveness [40, 42]. For the intercept term b in the Paired-Gaussian model, and all subsequent models discussed in this thesis, the `brms` approach for determining the initial value of the mean parameter is to use the median of the data. The scale parameter is the maximum of either the median absolute deviation (*mad*) of the data, or 2.5, and is also used to model the standard deviation of the Gaussian distribution.

Paired-Skew-Normal. As foreshadowed, the Skew-Normal is a generalization of the Gaussian distribution that allows for skewness. Algorithm 4.2 on the next page describes the Bayesian model used to estimate the location, scale, and skewness of the risk-adjusted score differences of pairs of systems. The weakly-informative prior for the skewness parameter λ

Algorithm 4.2: The Paired-Skew-Normal posterior log-density accumulator. Parameterizing the shape parameter based on the mean and standard deviation is derived from Bürkner [37].

```
Input: The array $\Delta$ of IR effectiveness scores differences of two systems, with
       current chain (or, initial) proposals for $\langle b, \sigma, \lambda \rangle$ used to explore the posterior
       distribution.
Output: The accumulated log-posterior density for the given inputs, for the MCMC
        sampler to probabilistically determine whether remaining in the current
        position or moving to this location is optimal.
                                   // Accumulate prior density using brms defaults
1  $location \leftarrow median(\Delta)$
2  $scale \leftarrow \max\{mad(\Delta), 2.5\}$
3  $lp \leftarrow student\_t\_lpdf(b, 3, location, scale)$
4  $lp \leftarrow lp + student\_t\_lpdf(\sigma, 3, 0, scale)$
5  $lp \leftarrow lp + normal\_lpdf(\lambda, 0, 4)$
                   // Parameterize the skewness in terms of mean and standard deviation
6  $\delta \leftarrow \lambda / \sqrt{1 + \lambda^2}$
7  $\omega \leftarrow \sigma / \sqrt{1 - 2/\pi \times \delta^2}$
8  $\xi \leftarrow b - \omega \times \delta \times \sqrt{2/\pi}$
                                                  // Accumulate likelihood density
9  for $y \in \Delta$ do
10     $lp \leftarrow lp + skew\_normal\_lpdf(y, \xi, \omega, \lambda)$
11 end
12 return $lp$
```

is a default zero-centered brms prior, specified as $\lambda \sim N(0, 4)$. Line 6 of Algorithm 4.2 enforces the joint relationship between the shape and the scale parameters of the Skew-Normal distribution (another pragmatic brms default).

4.1.2 Experimental Setup

To investigate the difference between Paired-Gaussian and Paired-Skew-Normal when modeling risk-adjusted score differences between pairs of systems, the Chapter 3 head-to-head scenario is used to explore their relative effectiveness. Recall that this scenario involved a comparison between the best submitted run on the Robust04 collection pircRB04td2, versus a BM25 baseline run with default parameters; scored using AP. As Sakai [159] explains that the ability to detect a significant difference is jointly related to effect size and sample size, the Robust04 topic-set is sub-sampled to the more typical IR collection size of 50 topics. (This approach was also used in Chapter 3.)

Bayesian MCMC. To fit either of the Paired-Gaussian or Paired-Skew-Normal models, the brms package is used, which provides a convenient interface via R to the Stan programming language. Stan is a probabilistic programming language that uses NUTS to perform MCMC sampling [101]. Recall from Figure 2.5 on page 44 that beyond specifying the model, the MCMC sampler requires information on how many chains to compute C, the number of iterations to run I, and the number of iterations to discard as warm-up W.

Hamra et al. [93, p. 629] mention that there are "no hard and fast rules" for choosing these parameters, and that the choice of C, I, and W is a "trial-and-error process". Rather than using different sampler properties for each model, the same values are used for all models in this chapter; derived from validating the most computationally intensive model Paired-Skew-Normal over multiple systems, defined in Section 4.3. The `brms` default approach for selecting the W number of burn-in iterations to discard is to remove the first 50% of the iterations. That leaves deciding on the number of chains C and iterations I:

- $C = 12$ chains are used, adhering to the Lambert [118, p. 314] recommendation that a "few tens of chains" be used for more complex models.

- $I = 12{,}000$ iterations are used, as this is the number of iterations (when the first $6{,}000$ burn-in iterations are filtered out) provides support for the effective sample size (n_{eff}) of many system group effects to exceed the $10{,}000$ value recommended by Kruschke [116] to support 95% credible intervals.

Although these settings are superfluous for the simpler paired models, they are completed in seconds on a modern consumer-grade laptop, and more samples provide more stable distributional parameter estimates than fewer samples. An additional benefit of using the `brms` package is that the computed model object can be inspected using the `shinystan` package in R, which provides many useful visualizations and diagnostics of the MCMC process. Quantitative posterior diagnostics are also tabulated in the results for chain convergence, and the effective sample size of each parameter via `shinystan`.

Risk Overlay. Recall from Chapter 3 that the direction of the URisk [207] statistic was inverted in URisk^{-} defined in Equation (3.2) on page 64, so that a higher score indicates greater risk. A URisk^{-} value greater than zero indicates that the challenger is riskier than the champion against the risk-level r.

The choice of risk-parameter is $r = 5$ for tabulated analysis, a moderate–high value from the set of conventional IR values: $1, 2, 5, 10$. These values are among the most common values used in the literature, as identified in Chapter 3 in Table 3.1 on page 64. The AP metric continues to be reported for all results, as it is the most common metric used in the literature, and was used in Chapter 3 for distributional analysis.

4.1.3 Experimental Analysis

Table 4.1 on the next page shows the posterior summary statistics exported from the diagnostic tool `shinystan` implemented in R, which tabulates the results of four models, each built using either the Paired-Gaussian or Paired-Skew-Normal with no risk transformation applied or with $r = 5$. The left-most column describes the parameter that has been estimated, where

Parameter	$\hat{R}$	n_{eff}	Mean	SD	Q2.5	Median	Q97.5
Paired-Gaussian ($r = 1$)							
Location (μ)	1.00	34,204	-0.12	0.02	-0.16	-0.12	-0.09
Scale (σ)	1.00	32,961	0.13	0.01	0.11	0.13	0.16
Paired-Skew-Normal ($r = 1$)							
Location (ξ)	1.00	34,931	-0.12	0.02	-0.16	-0.12	-0.09
Scale (ω)	1.00	32,087	0.13	0.01	0.11	0.13	0.17
Shape (λ)	1.00	25,768	1.04	1.67	-1.70	0.94	4.87
Paired-Gaussian ($r = 5$)							
Location (μ)	1.00	34,755	-0.08	0.04	-0.17	-0.08	0.00
Scale (σ)	1.00	34,494	0.31	0.03	0.25	0.31	0.38
Paired-Skew-Normal ($r = 5$)							
Location (ξ)	1.00	23,815	-0.03	0.03	-0.09	-0.03	0.04
Scale (ω)	1.00	24,489	0.25	0.03	0.20	0.24	0.31
Shape (λ)	1.00	23,660	6.67	2.13	3.25	6.41	11.46

Table 4.1: Posterior summary statistics for the Bayesian models computing either the Paired-Gaussian or Paired-Skew-Normal for $r = 1$ (no risk) and $r = 5$ (losses in the challenging retrieval model counting five times as much) for the AP measure on a head-to-head comparison of a BM25 run against the best submitted run pircRB04td2 on the Robust04 track. The shape parameter λ of the Skew-Normal distribution grows as r increases, and consequentially has a different location value compared to the Paired-Gaussian on $r = 5$.

the summary columns on the right of the table describe the spread of each parameter over the posterior distribution. The second and third columns provide diagnostic statistics for the MCMC process over each model. Respectively, no $\hat{R}$ values above 1.0 are reported, therefore the MCMC chains have mixed well for each model. All effective sample sizes $n_{eff} > 10{,}000$, which is the recommended threshold for credible intervals to be 95% accurate in the tails of each parameter distribution [116].

After confirming that the posterior distribution of each parameter is valid, the parameters can be compared between each modeling approach. Highlighted in Table 4.1 is are the shape values λ for each of the Paired-Skew-Normal models, computed over either no risk $r = 1$, or with risk $r = 5$. Note that when no risk is applied, the shape parameter is 1.04, which implies the distribution is not perfectly symmetric, but is close to a normal distribution. When risk is applied, the shape parameter is 6.67, which is much higher than when no risk is applied. That implies that the mean score distribution does get more skewed when risk is applied, providing further support for the results of Chapter 3. When exploring what the practical implication of modeling the risk with a skewness parameter, observe the difference in location values between the Paired-Gaussian and Paired-Skew-Normal pairs for each risk level. However, when there is no risk transformation, the location parameters between

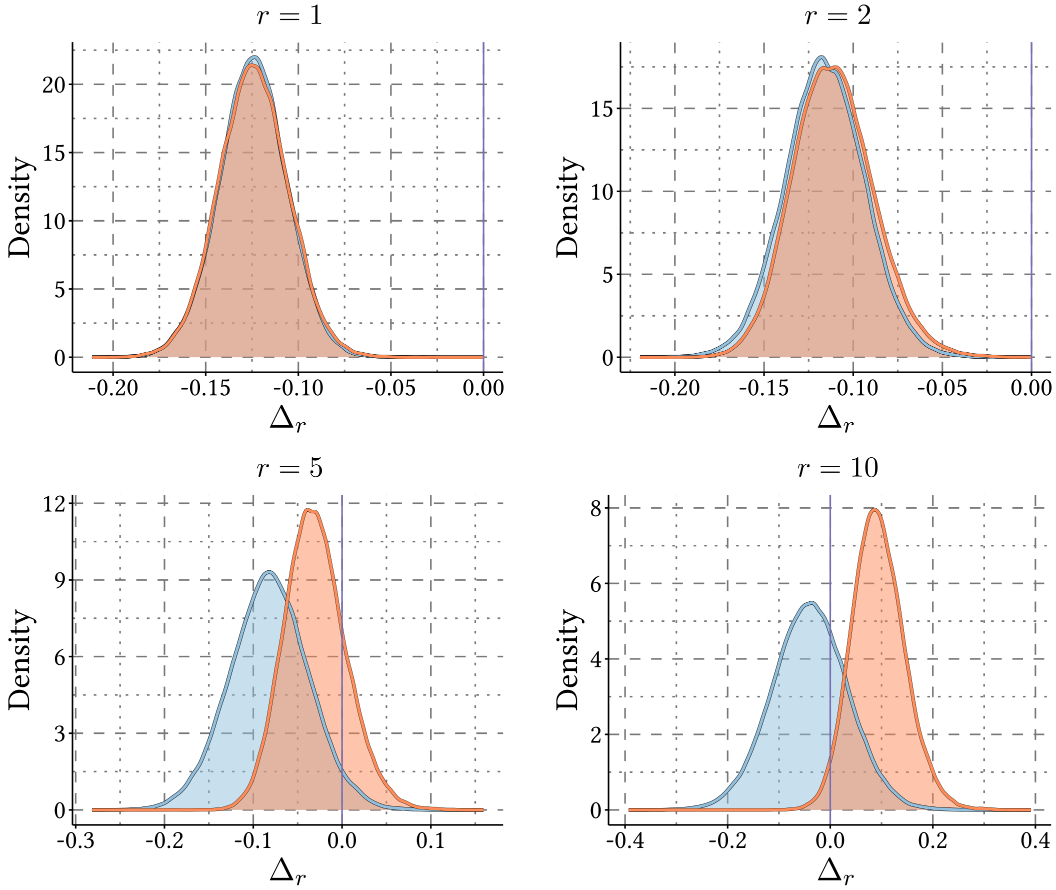


Figure 4.2: Estimates of the location parameter for paired risk-adjusted score differences when a Paired-Skew-Normal model (in orange) is compared against a Paired-Gaussian (in blue). Each risk-adjusted score is against a BM25 champion run on the Robust04 collection compared to the best run `pircRB04td2` on the AP metric, where the topics have been sub-sampled to 50 to emulate a more traditional IR collection size.

the two models are close for two decimal places. Conversely, when risk is applied, the location values are much more different, with the Paired-Skew-Normal providing a larger location value, indicating that swapping the champion system with the challenger is more likely to be riskier than the Paired-Gaussian.

To aid in visualizing the difference between the Paired-Gaussian and Paired-Skew-Normal models, Figure 4.2 presents the $r \in \{1, 2, 5, 10\}$ density plots of the location parameter for each posterior when risk-adjusted score differences are modeled. These plots further confirm that the location distribution becomes one-sided as r increases: the difference at $r = 1$ and $r = 2$ is negligible; but for $r = 5$ and $r = 10$ the intervals shifts to the right. The Skew-Normal distribution is produces a tighter credibility interval for the μ parameter than the Gaussian alternative after the score differences have been adjusted to the $r = 5$ and $r = 10$ levels.

4.1.4 Outcomes

This section has presented a novel methodology for Bayesian modeling of paired risk-adjusted score differences, finding that the Paired-Skew-Normal distribution is useful in providing more sensitive comparisons of location parameters for inferential purposes; answering S-RQ4.1. As Chapter 3 exhaustively finds that the skewness of risk-adjusted score distributions is not reliant on any particular run or corpus combination, this chapter uses a straightforward paired comparison to both validate the utility of the model, as well as to better understand how much more power is gained by using the Paired-Skew-Normal Bayesian inferential approach over nonparametric tests. Recall that in Chapter 3 in Figure 3.9 on page 87, nonparametric testing resulted in wider confidence intervals than the parametric t-distribution on average. This chapter has shown that the Skew-Normal distribution provides tighter confidence intervals than the Gaussian alternative. Although this method provides another tool for the IR community to use when comparing the effectiveness of pairs of systems, the key goal of this thesis is to evolve this paradigm to support multiple system comparisons. The Paired-Skew-Normal approach is used as a stepping stone into achieving that goal later into the chapter in Section 4.3. However, in order to make use of what has been learned, we must first understand how to perform multiple comparisons in the Bayesian paradigm, an unexplored problem in IR which is the focus of the next section.

4.2 Bayesian Inference for Multiple System Comparisons

Often there is more than one challenger system to be compared to a champion system. Section 4.1 explored modeling risk-adjusted scores between two systems using a Paired-Skew-Normal. The results of that experiment has provided further evidence of the Chapter 3 observation that modeling shape is an important factor for risk scores. In order to extend the Section 4.1 findings to model risk-adjusted scores over many systems, this section explores a Bayesian methodology for modeling standard IR effectiveness scores over many systems; to be extended to risk-adjusted scores in Section 4.3.

This section answers S-RQ4.2: *How does using previous system artifacts affect Bayesian inferential results for IR test collections?* First, the Bayesian hierarchical modeling methodology employed in Bayesian inferential analyses involving multiple groups is defined in Section 4.2.1 on the following page. Then, a small-scale example is presented in Section 4.2.2 commencing on page 107, which describes the methodology for interpreting statistical significance in the Bayesian setting. Section 4.2.3 on page 110 describes the methodology used to expand the analysis to complete datasets, where Section 4.2.4 starting on page 112 demonstrates the statistical power achievable when using Bayesian models with full datasets. Section 4.2.5 on page 113 shows the effect on the outcomes of significance when varying counts of artifact systems are supplied to the model. Finally, Section 4.2.6 commencing on page 117 answers the research questions and provides commentary on some practical use-cases of Bayesian modeling of benefit to IR practitioners.

4.2.1 Theory

The novel proposal for modeling IR effectiveness scores across many systems is now described. *Bayesian hierarchical modeling* (BHM) is a little-known, but key concept in Bayesian inference that pulls estimates of an individual group effect towards the average group effect, also known as *partial pooling*, and which Gelman et al. [88] notes is the Bayesian approach for multiple comparison correction. Since other IR tests assume normality, a simple Gaussian model is used to describe these scores, as was done in the Paired-Gaussian approach explored in Section 4.1 that investigated paired score distributions.

The BHM-Gaussian model assumes that the distribution of effectiveness values is normally distributed across the matrix of systems and topics[2]:

$$\begin{aligned} y_{ij} &\sim N(\hat{\alpha}_i + \hat{\beta}_j, \sigma^2) & b &\sim t(3, \mathit{median}, \max\{\mathit{mad}, 2.5\}) \\ \hat{\alpha}_i &= \omega_{\alpha,\alpha_i}\mu_\alpha + (1 - \omega_{\alpha,\alpha_i})\alpha_i & \mu_\alpha, \mu_\beta, \sigma &\sim t(3, 0, \max\{\mathit{mad}, 2.5\}) \\ \hat{\beta}_j &= \omega_{\beta,\beta_j}\mu_\beta + (1 - \omega_{\beta,\beta_j})\beta_j & \alpha_i, \beta_j &\sim N(0, 1)\,, \end{aligned} \tag{4.1}$$

where y_{ij} is an effectiveness score parameterized by topic j and system i. The topic and system effects, β_j and α_i respectively, are moderated by *partial pooling* in the corresponding $\hat{\beta}_j$ and $\hat{\alpha}_i$ [84], where $\omega_{Y,y}$ is the pooling factor that measures the simulated strength of the hypothetical population Y versus the observed group effect y:

$$\omega_{Y,y} = 1 - \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_y^2}\,. \tag{4.2}$$

The expression $y_{ij} \sim N(\hat{\alpha}_i + \hat{\beta}_j, \sigma^2)$ is a conventional shorthand in statistics that declares y_{ij} to be a normally distributed variable. The expanded form is a linear equation:

$$y_{ijk} = \mu + \hat{\alpha}_i + \hat{\beta}_j + b\,, \tag{4.3}$$

which is similar in form by means of the pooling factor transformation to the ANOVA Equation (2.33) on page 42 without interaction effects, including a correction similar to Tukey's HSD defined in Equation (2.30) on page 41.

The standard deviation σ of various parameters in the model is specified with a three-parameter Student t-distribution prior, where its arguments correspond to the non-informative defaults in `brms` for the Gaussian distribution (discussed in Section 4.1). The above approach is an extension of the original *Model 2* specified in Carterette [42], with marginally more informative priors than the Jeffreys' prior:

$$\sigma \sim \log\left(1/\sigma\right)\,. \tag{4.4}$$

[2] This amends Equation 3 of the SIGIR paper [25], which omitted the partial pooling notation, which was addressed and accepted in the ECIR [26] work following this chapter.

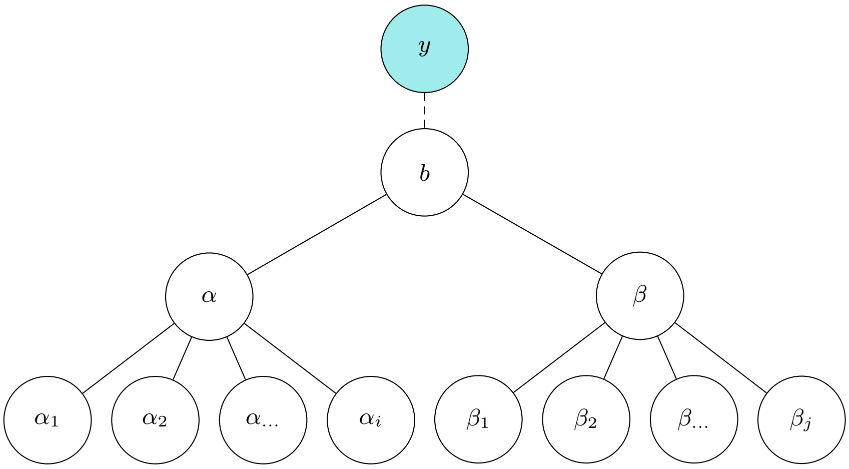


Figure 4.3: The hierarchy of system and topic effects investigated in this work, based on the explanation of Bayesian hierarchical modeling presented in Dietze [64]. Each individual system (α) and topic (β) effect on the leaf-nodes of the tree is moderated by the average effect over all systems or topics (partial pooling), where any error that cannot be ascribed to these effects is propagated into an error term.

Figure 4.3 conceptualizes the system and topic effects into a hierarchical model by displaying a tree diagram. When the *general linear mixed model* (gLMM) is specified, the individual contributions of independent systems and topics is modeled, however, each system and topic effect is moderated by the average effect over all systems or topics, respectively. When an effect cannot be ascribed to the system or topic, it is propagated into an error term. The presented tree diagram defines a hierarchical model using partial pooling as described previously.

Figure 4.3 also helps to understand what is meant by partial-pooling. If a model were to use full-pooling, a average α would represent any system, and there would be no leaf-nodes from α (the same applies to β). In practice, full-pooling cannot be used to infer which system is better or worse than another. As for the other extreme, no pooling would branch each system and topic effect from the error term, providing no ability to correct for multiple comparisons.

The BHM-Gaussian model is now formally defined in Algorithm 4.3 on the following page. The highlighted code-blocks and inputs outline the extensions to the Paired-Gaussian approach shown in Algorithm 4.1 on page 99 which are necessary to implement Bayesian hierarchical modeling. Much of the additional code is related to specifying the new mean parameter for the normal distribution based on all of the systems and topics, honoring the form of Equation (4.3).

Factor Analysis. The BHM-Gaussian approach describes an effectiveness score as a linear combination of system and topic effects. However, a range of recent results in regard to the variance of IR effectiveness scores contribute to understanding how to reduce error when modeling scores. Ferro and Silvello [74] use a general linear mixed model (gLMM) to approx-

Algorithm 4.3: The BHM-Gaussian posterior log-density accumulator. The highlighted parts describe the added components to the Paired-Gaussian approach to make it a hierarchical model.

```
Input: The array $S$ of IR effectiveness scores of many systems, with current chain
       (or, initial) proposals for $\langle b, \sigma, \mu_\alpha, \alpha_i, \mu_\beta, \beta_j \rangle$ used to explore the posterior
       distribution.
Output: The accumulated log-posterior density for the given inputs, for the MCMC
        sampler to probabilistically determine whether remaining in the current
        position or moving to this location is optimal.
                                        // Accumulate prior density using brms defaults
$location \leftarrow median(S)$
$scale \leftarrow \max\{mad(S), 2.5\}$
$lp \leftarrow student\_t\_lpdf(b, 3, location, scale)$
$lp \leftarrow lp + student\_t\_lpdf(\sigma, 3, 0, scale)$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\alpha, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\alpha_i, 0, 1)$                // Standard normal prior for $\alpha_i$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\beta, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\beta_j, 0, 1)$                 // Standard normal prior for $\beta_j$
                                                   // Accumulate likelihood density
for $y_{ij} \in S$ do
    $\hat{\alpha_i} \leftarrow \mu_\alpha \times \alpha_i$      // Model interaction of current system and average system
    $\hat{\beta_j} \leftarrow \mu_\beta \times \beta_j$         // Model interaction of current topic and average topic
    $\mu_{ij} \leftarrow b + \hat{\alpha_i} + \hat{\beta_j}$    // Set linear predictors
    $lp \leftarrow lp + normal\_lpdf(y_{ij}, \mu_{ij}, \sigma)$  // Probability density function
end
return $lp$
```

imate system performance as an amalgamation of system effect, topic effect, and system-topic interactions, establishing the relative impact on retrieval scores of stop lists, stemmers, and retrieval models. The gLMM and two-way ANOVA approaches were also used to explore shard and topic-shard interaction effects in distributed retrieval, greatly reducing the regression error [75]. Carterette [40] compares a Bayesian linear model against the t-test, noting that using a t-test means implicitly accepting a model too.

It is important to recognize that Bayesian simulation is substantially more expensive than classical statistics, as Bayesian simulations find many thousands of credible parameters to estimate a model, whereas classical statistics provide single point-wise parameter estimates. Those component effects are an interesting point to consider in future work to improve the reliability of score point estimates and credible intervals, especially if MCMC simulation efficiency and computing hardware improves.

4.2.2 Small-Scale Example

Figure 4.4 on the next page provides an overview of the experimental scenario of this chapter from here in. The key idea is to explore the Bayesian hierarchical modeling methodology to perform statistical inference on the observed IR rankers in one go, rather than running many

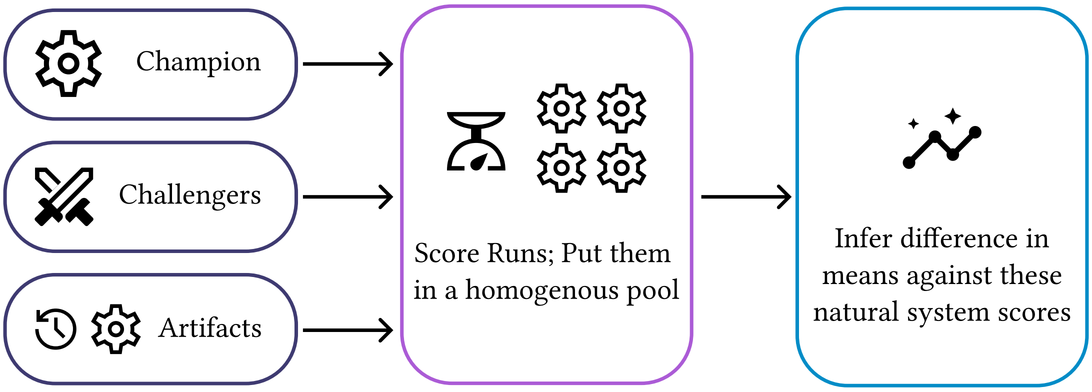

Champion
Challengers
Artifacts
Score Runs; Put them in a homogenous pool
Infer difference in means against these natural system scores

Figure 4.4: A high-level overview of the inferential evaluation scenario. Information from previous systems (artifacts) is combined with runs, to determine whether to switch from the champion system to any one of a set of challenger systems.

		Topics					
System	Measure	301	306	311	316	321	Mean
Champion	AP	0.055	0.209	0.484	0.621	0.286	0.331
Challenger 1	AP	0.062	0.243	0.423	0.620	0.340	0.338
Challenger 2	AP	0.061	0.243	0.433	0.619	0.344	0.340
Challenger 3	AP	0.045	0.189	0.460	0.617	0.299	0.322
Challenger 4	AP	0.189	0.087	0.323	0.651	0.339	0.318
Artifacts	Mean	0.105	0.133	0.246	0.512	0.168	–
Artifacts	Median	0.066	0.112	0.212	0.599	0.159	–
Artifacts	IQR	0.143	0.132	0.341	0.217	0.151	–

Table 4.2: System-topic scores used to demonstrate the utility of the Bayesian hierarchical inference model. Artifacts correspond to summary statistics for the pool of 79 artifact systems used to build the posterior.

null hypothesis tests at a time. To explore that goal, Table 4.2 provides a cutback example of real system-topic scores over five systems and topics for illustrative purposes, with a champion system compared against four challengers. Corresponding to the artifacts component of Table 4.2 in the bottom-left, a sample of 79 artifact systems (not shown) over the same test data is used to generate a posterior distribution. These values are subsets of real AP score data described later in the chapter.

Figure 4.5 on the following page shows the credible intervals of model variables when modeling the Table 4.2 data using the BHM-Gaussian approach after MCMC. In addition to the credible intervals, each observation has an associated predictive interval generated out of the simulation at the bottom of the figure. With only five different scores associated with each system, it makes sense that no clear winner (nor loser) emerges. The context provided by the artifact systems makes it apparent that all of the observed scores are in line with what

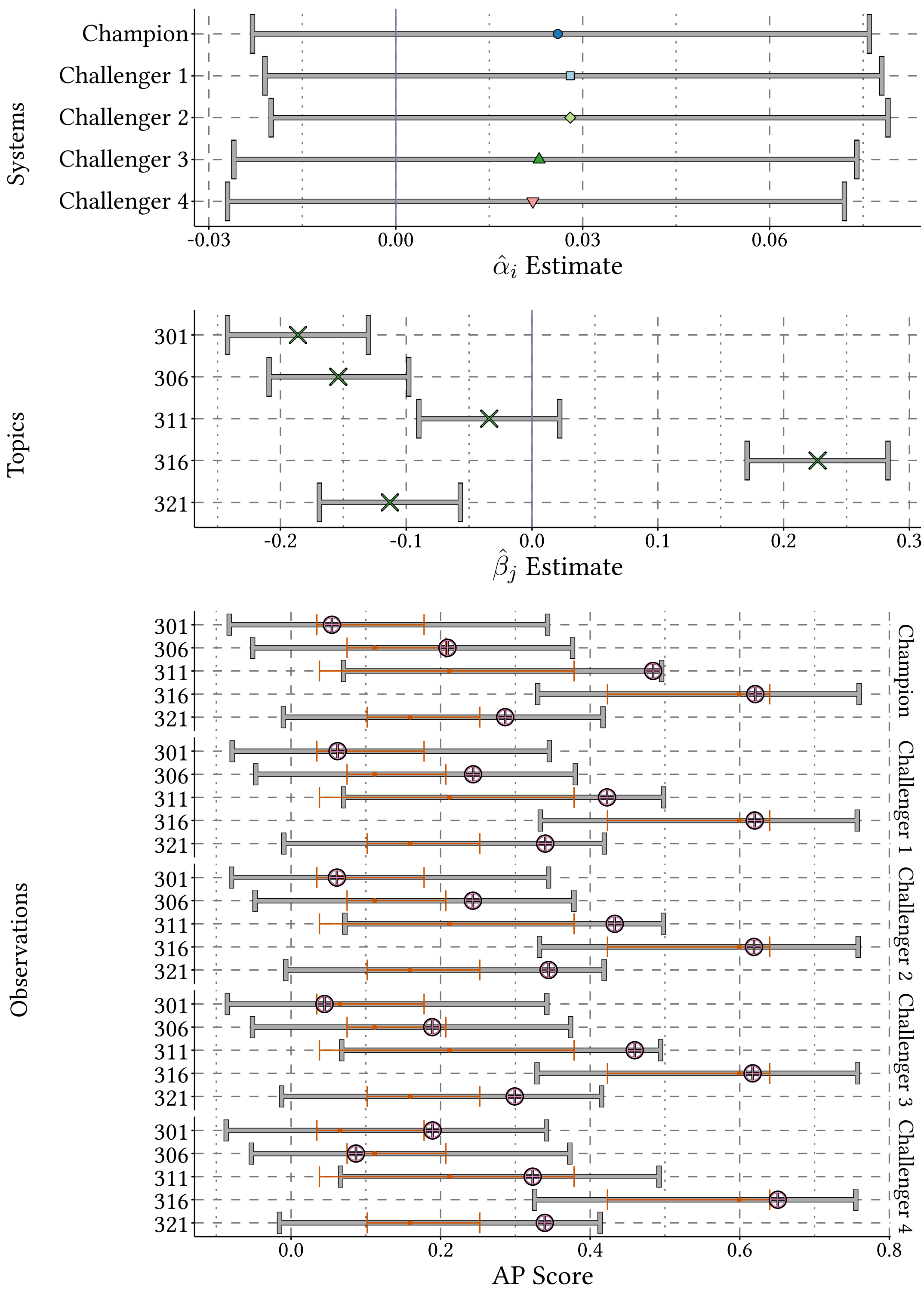


Figure 4.5: The information made available for BHM inferences. At top, credible intervals for the system effect intercepts are plotted with a point estimate of the random effect; in the middle topic effect estimates are shown in green crosses with their corresponding credibility intervals. Forecasts for 95% intervals based on the score observations are shown with a thick white error-bar in the bottom part of the diagram, with plus symbols corresponding to actual scores (Table 4.2). Embedded within the error-bar is the median and interquartile range in orange for the artifact systems.

Collection	Citation	Documents	Unique Terms	Total Terms	Topics
Robust04	[199]	528,155	664,603	253,367,449	250
TREC17	[13]	1,855,658	2,970,013	1,285,653,766	50
TREC18	[3]	595,037	1,478,198	481,432,022	50

Table 4.3: Statistics for the collections used in Sections 4.2 and 4.3.

is expected. It is apparent that even with this small dataset where only five systems are considered, that topic 316 is significantly *easier* than all others perform well on AP, and topic 311 is significantly easier than topics 301 and 306 on a 95% credible interval. That does not imply that topic 316 will always be significantly easier than the other topics considered. Rather, with the limited knowledge at hand, the framework enables inferential assertions, where further studies provide more evidence to support or refute claims.

4.2.3 Experimental Setup

Section 4.2.2 highlighted the potential of BHM-Gaussian to provide more advanced statistical inferences with more information than the classical null hypothesis testing paradigm. With that, the next step is to explore the utility of the approach on real data at the expected scale. Although the runs used in this experiment were not pooled, the artifact systems used in the experiments were assessed when forming the relevance judgments, and so, the inferences are further derived with respect to these systems.

Suppose an organization provides a search feature that ranks newswire documents, and they wish to improve the user experience by improving the search effectiveness. The currently implemented retrieval model is BM25 with default Terrier parameters, and three teams have been charged independently with the task of proposing a challenger system to become its replacement. As A/B testing can expose users to a new system that is worse than the current system, the organization wishes to verify that the challenger system poses no significant chance of harm before deploying it as an A/B candidate.

Corpora. Three collections of news documents are employed: the classic Robust04 corpus, and the more recent *New York Times* and *Washington Post* collections, TREC17 and TREC18 respectively. Table 4.3 provides collection statistics on the above corpora used in this study.

- Robust04 [199]: As Robust04 has 250 topics (technically 249, as topic 672 did not receive judgments), and sample size has a direct bearing on significance test outcomes, two topic-set sizes of Robust04 are explored; the full topic set, and every fifth topic (301, 306, and so on) to obtain a 50 topic sub-sample. The collection has $\approx$ 528k documents, $\approx$ 664k unique terms, and $\approx$ 253M terms in total. Of the 110 runs submitted, $m = 79$ are available as artifacts after the bottom 25% had been removed, and the top-three removed to create Challenger 4.

		AP Score		
System	Description	Robust04	TREC17	TREC18
Champion	BM25	0.274	0.210	0.236
Challenger 1	DPH + Bo1	0.323	0.289	0.301
Challenger 2	DPH + DFR + SD +Bo1	0.322	0.290	0.300
Challenger 3	DPH + DFR + SD	0.264	0.216	0.231
Challenger 4	Top-3 TREC Runs Fused	0.380	0.572	0.459

Table 4.4: Hypothetical experimental systems with their AP effectiveness, used to explore the ability to perform inference using BHM on the scenario set-out in Figure 4.4 on page 108.

- TREC17 [13]: The TREC CORE 2017 Track exercise, with ≈ 1.85M documents, ≈ 2.97M unique terms, and ≈ 1.28T terms in total. There were 75 runs submitted, with $m = 53$ after removing the bottom 25% of runs and the top-three.
- TREC18 [3]: The TREC CORE 2018 Track exercise, ≈ 595k documents, ≈ 1.47M unique terms, and ≈ 481B terms in total. Of 72 runs submitted, $m = 51$ are used as artifacts.

Runs. The champion and first three challengers are from the Terrier v5.2 search engine, taking configurations originally proposed in the *SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR)* workshop [14] as representative retrieval models. Runs are scored using AP, the official metric for all three of the test collections explored in the study. Table 4.4 describes these runs, and their corresponding mean-over-topics AP effectiveness computed on three collections[3]. Each run uses a hypergeometric weighting model DPH, with various settings with respect to divergence from randomness, term proximity, and query expansion. For Bo1 query expansion, 10 feedback documents with 25 expansion terms are used.

The fourth challenger was selected as it is known to be highly effective, formed by fusing the best three participant runs submitted to the TREC rounds associated with the three document collections. On average, all challenger runs have a greater mean effectiveness than the BM25 run on all three collections with the exception of Challenger 3; and the control run Challenger 4 that was intentionally selected to be effective, has a large mean AP score improvement of $+0.110$ on the champion system. These nine runs (three per collection) are omitted from the set of artifact systems used to compute the Bayesian posterior. To create Challenger 4 for Robust04, CombSUM [77] fusion was used; and for TREC17 and TREC18, reciprocal rank fusion [59] was applied. Both fusion techniques are simple and known to generate effective outcomes. All of the five runs are interesting in a risk-sensitive retrieval context, as query expansion is known to be volatile due to query drift, and rank fusion has been shown to reduce risk in prior work [22].

[3] See *Terrier v5.2 Documentation* for details of these options and settings, `https://github.com/terrier-org/terrier-core/blob/5.x/doc/index.md`, accessed on 17th May 2021.

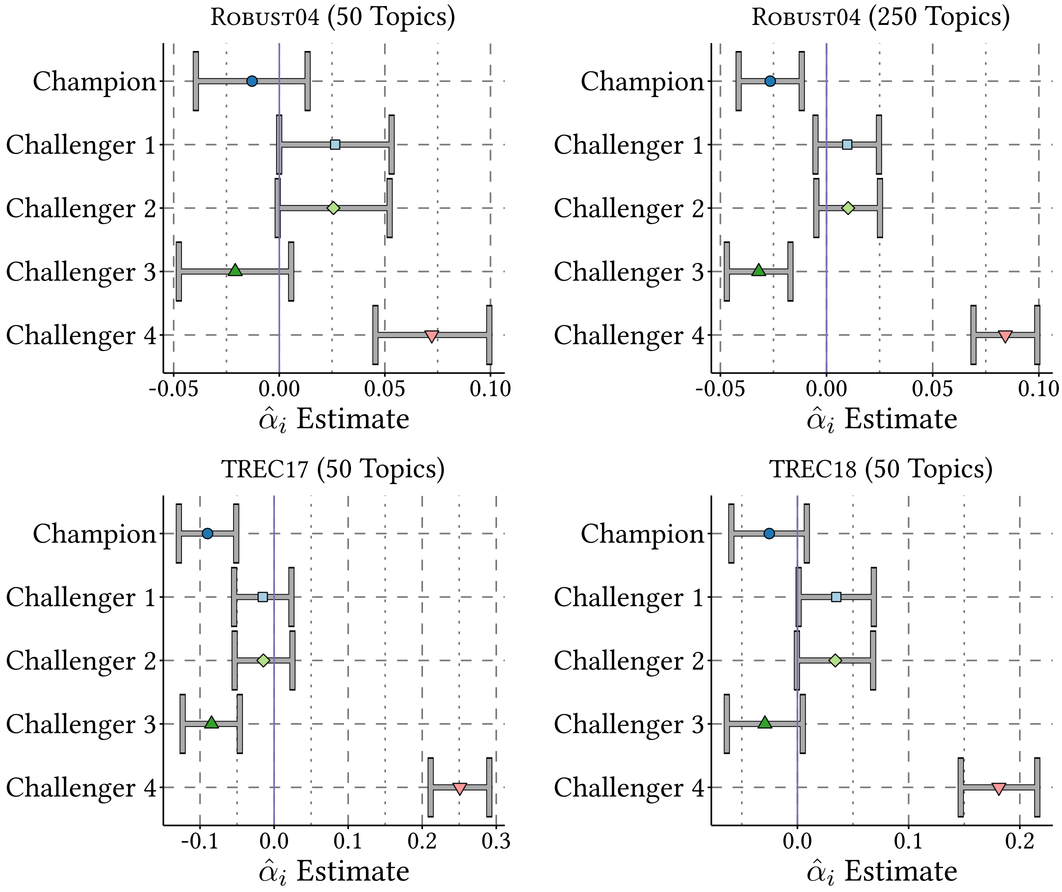


Figure 4.6: Parameter estimates of each of the systems listed in Table 4.4 on the previous page when a Bayesian hierarchical model assumes a Gaussian score distribution. The top row corresponds to Robust04 estimates on 50 and 250 topics, while the bottom row specifies two TREC CORE collections, the 2017 and 2018 tracks respectively.

Bayesian MCMC. As mentioned in Section 4.1.2 on page 100, the MCMC simulations for each dataset are executed using 12 Markov chains for 12,000 iterations, where the first 6,000 iterations of warm-up are dropped, informed by post-hoc diagnostics. All BHM MCMC is computed with `brms`, a framework with intelligent generalized defaults for hierarchical modeling in R. From here-on, all results involving MCMC simulations have been validated to have converged, and the effective sample size of all parameters $n_{\mathit{eff}} > 10{,}000$. The experiments reported here are the first to use BHM to allow inference on the effectiveness of multiple IR retrieval systems.

4.2.4 Results

After simulating a BHM model including artifacts, challengers, and the champion system, Figure 4.6 plots each of the system parameter estimates in the model for each collection. The top row presents the credible intervals for the Robust04 collection on 50 and 250 topics, while the bottom row shows the TREC CORE 2017 and 2018 collections. Zero on the horizontal axis (the $\hat{\alpha}$ estimate) corresponds to the average system effect. When 50 topics are used on all three collections, only Challenger 4 is statistically significantly more effective than the BM25 champion system. But, the BM25 champion is on the cusp of significance against Challengers 1 and 2 on the CORE collections. In the top-right quadrant, as the Robust04 topic set grows from 50 to 250, Challengers 1 and 2 are now significant.

Figure 4.7 on the next page shows the relative topic difficulty over 95% credible intervals for topic scores. Mizzaro and Robertson [133] proposed *average average precision* (AAP) as a measure of topic difficulty over many systems. As a byproduct of modeling the topic effect hierarchically, the MCMC process enables making distributional comparisons of topic difficulty. This incidentally forms a methodology for inferentially comparing topic difficulty with multiple comparison correction applied in this context too. In interpreting the $\hat{\beta}_j$ weights for each topic, a positive value indicates that the topic is easier (on average, across the topics considered) for systems to retrieve relevant documents on, while a negative value indicates that the topic is more difficult. Where the goal of a practitioner may be to trichotomize topics into easy, medium, and hard, $\hat{\beta}_j$ estimates cutting across the 0 value may be considered medium difficulty, as 0 corresponds to the mean topic difficulty across all topics. Credible intervals of $\hat{\beta}_j$ that do not include zero could be considered as easy or hard topics, depending on the sign of the estimate. When 250 topics are considered on the Robust04 collection, the credible intervals for the topic effects tighten, allowing for further significant differences to be observed between topics.

The BHM-Gaussian approach has identified significant differences between systems and topics in the above inferential scenario involving a champion, challenger systems, and artifact systems as background information. This partially answers S-RQ4.2, in the situation where all artifact systems are available on a collection, significance between systems is more readily achieved. When 50 topics are used on Robust04 the Champion system is significantly different to Challenger 4. Further statistical power is achieved when the full 250 set is used on Robust04: Challengers 1, 2, and 4 are significant against the Champion. To more completely answer S-RQ4.2, the effect of the count of artifact systems included in the sampling process and its relationship with achieving statistical significance now explored.

4.2.5 The Effect of Artifact Systems

The ability for the BHM-Gaussian method to detect significant differences between systems may be predicated on the count and effectiveness of the artifacts included in the model. This subsection explores that possibility. Since both the number of systems included in the model, as well as the effectiveness of the artifacts used, are likely to be important factors in how significance is determined using the BHM-Gaussian method, the count of artifacts m is varied in steps of $\{1, 5, 10, 20, 40\}$. Recall that the bottom 25% systems had been removed from the artifact pool prior to running any experiments, and the m artifacts are selected from the static set of the top 75% of runs on the collection. To observe the influence of retrieval effectiveness, these steps of systems are included in the model by best and worst AP effectiveness, to better understand how artifact systems influence the outcomes of the experiments. That is, the total

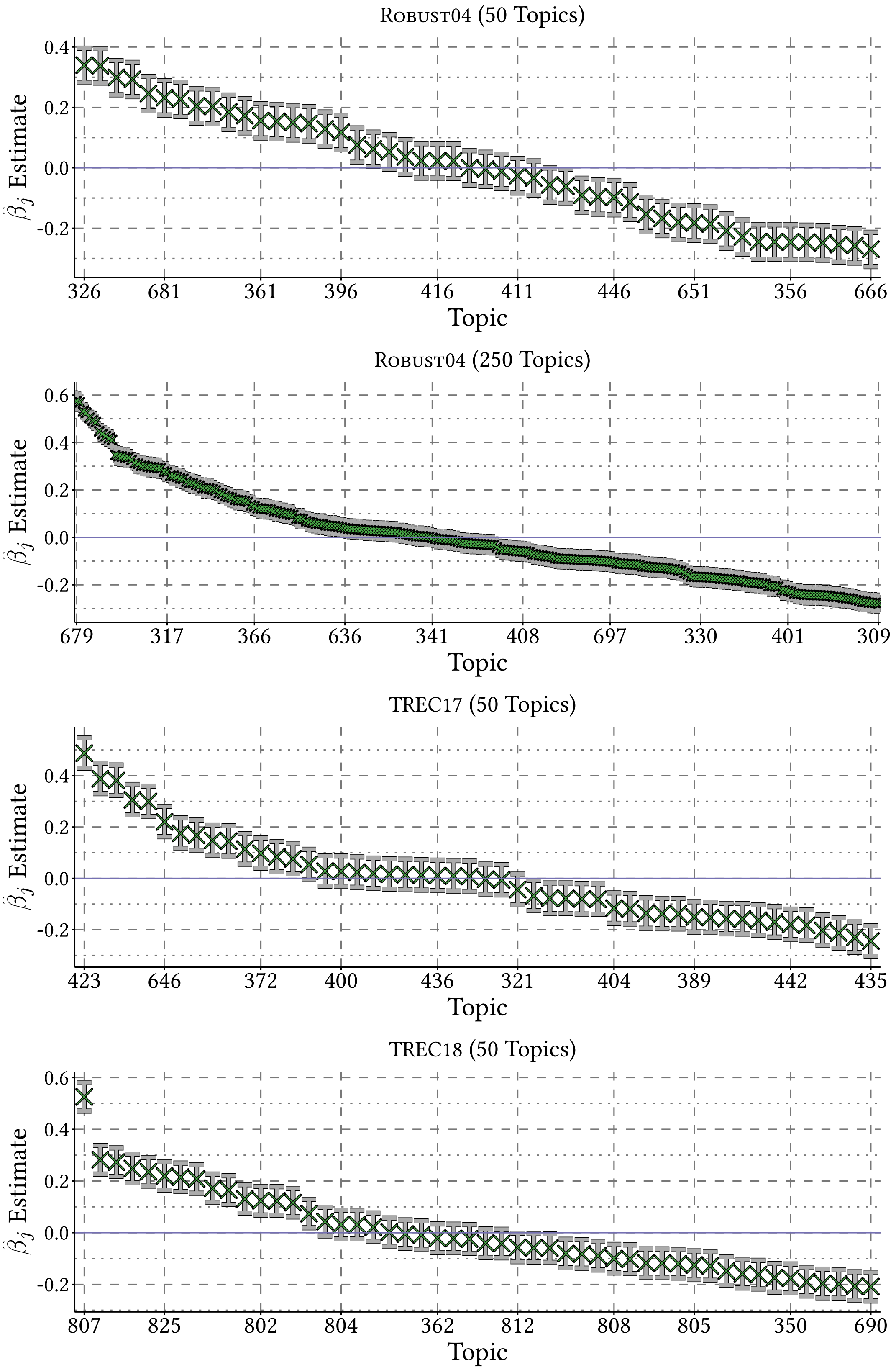


Figure 4.7: Per-topic parameter estimates for all systems in the evaluation pool (the systems listed in Table 4.4 on page 111 together with artifact systems) using a Bayesian hierarchical model that assumes a Gaussian score distribution.

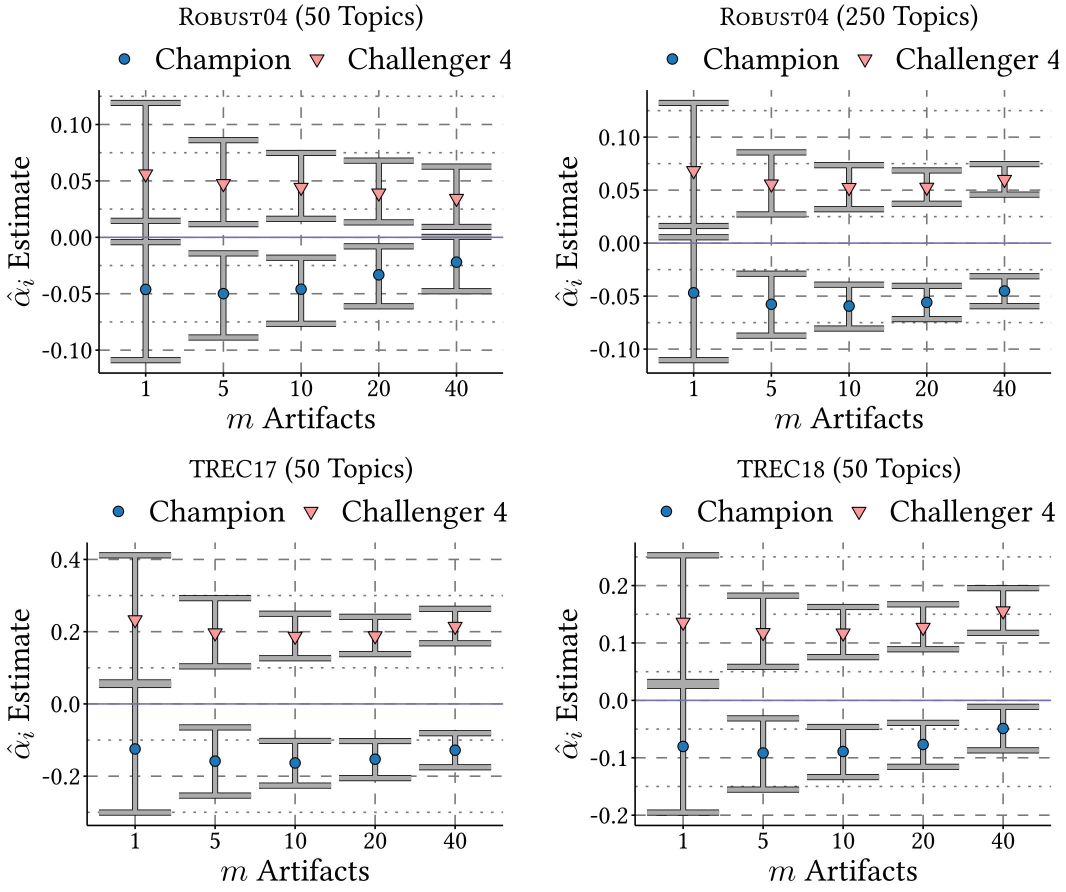


Figure 4.8: Exploring the ability to detect significance of a champion BM25 run against a high quality fusion run of the top-3 (held-out) best systems submitted to each evaluation campaign, when the number of artifact systems increase in decreasing effectiveness order (from most effective, downward).

count of systems provided to the BHM-Gaussian model in all cases is $m + 5$ to include the champion vs. challengers, with m varying as the set of artifacts is extended. The best-systems-first ordering is motivated by Armstrong et al. [15], which highlighted the importance of evaluating against strong baselines, and the worst-systems-first for contrast. Ideally, for a given m artifact pool, the BM25 Champion and fusion run Challenger 4 should be significantly different, and this is used as a reference point for answering S-RQ4.2.

Including Artifacts By Best Effectiveness. Figure 4.8 displays the credible intervals between the champion BM25 and Challenger 4 systems for the varied counts artifacts included in the model for the Robust04 collection by most AP effectiveness to least. When only one artifact is included, there is not enough information about the system effect to conclude whether the two runs are significantly different. Recall from the Figure 4.5 on page 109 example involving a small-scale experiment that the credible intervals overlap when not enough information is available. However, in every dataset, when the count of artifacts equals or exceeds 5, the

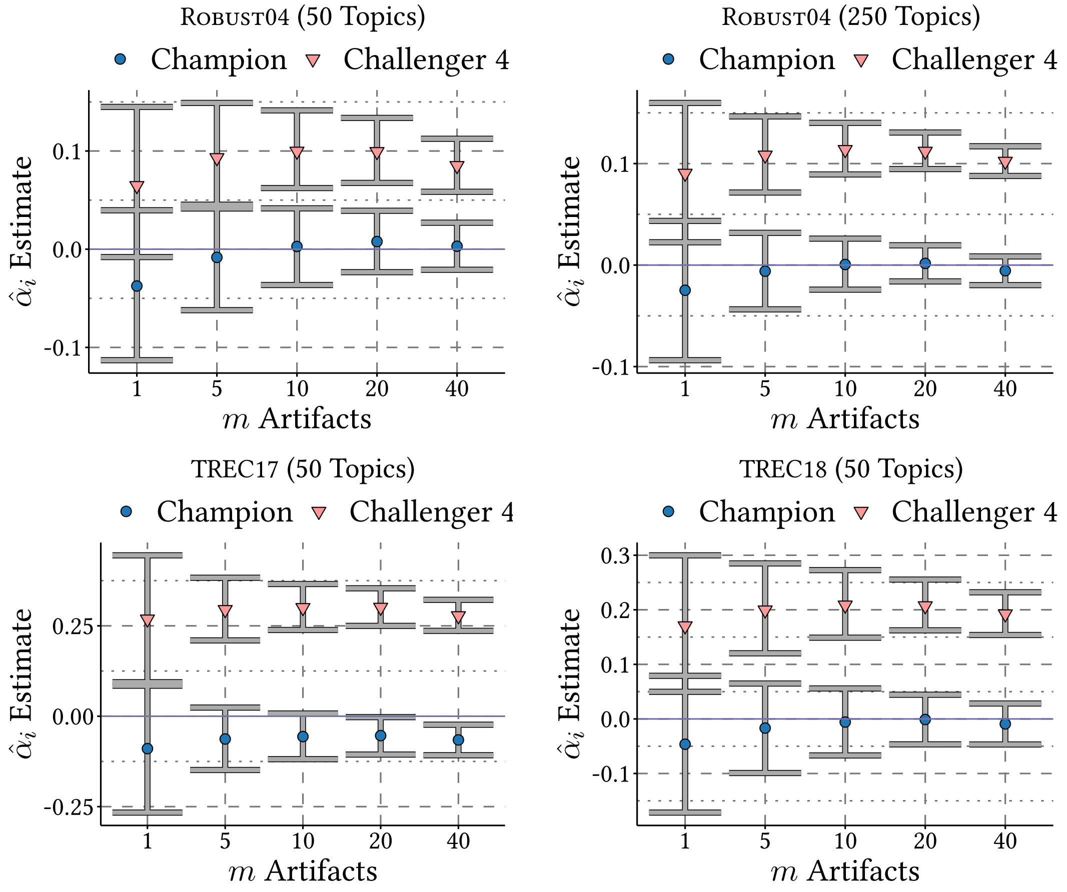


Figure 4.9: A graphic with similar parameters to Figure 4.8 on the previous page, except in this case, the worst systems are being gradually included into the artifact pool. Note that these poor performance systems are in the filtered top-75% of the original submissions to to avoid skewing results with erroneous systems.

BHM-Gaussian model is able to separate the two runs. When comparing between the 50 and 250 topic set sizes for Robust04, the separation between each of the runs is more pronounced on the 250 topic set for the same count of artifacts.

Interestingly, as the count of artifacts included increases on the 50-topic Robust04 dataset, the strength of the significance between the Champion and Challenger 4 systems did not increase, as was seen on all other datasets. That exception suggests that statistical significance cannot be expected to monotonically increase with respect to m. This outcome may be a consequence of the Robust04 collection being composed of a biased sample of difficult topics, where the topic effect grows to be more dominant than the system effect when more data is made available. When the full Robust04 topic set is considered however, the pattern of the system effects follows the CORE collection behavior. The case where artifacts are included by worst effectiveness is explored next to see if the same pattern is observed.

Including Artifacts By Worst Effectiveness. The same experiment is repeated, but this time the artifact systems are included in the model by worst effectiveness to best. Figure 4.9 on the previous page shows the outcome of the Champion system vs. Challenger 4 comparisons. Despite the location and width of the error bars being different for the reverse-ordered artifact inclusions, the outcomes of significance are similar between the best vs. worst effectiveness orders. The only difference in significance observed between the two artifact inclusion orderings is on the Robust04 50 topic set when 5 artifacts are included. In this case, the best-first ordering of the artifacts reports significance while worst-first does not.

4.2.6 Discussion

This section empirically evaluated a Bayesian methodology for detecting statistical significance in a multiple system comparison scenario. The BHM-Gaussian method was shown to be able to detect significant differences between systems in a variety of settings. An important factor in the ability to detect significance is the count of artifact systems included in the model. The outcomes of experiments exploring the contribution of effectiveness and count of artifact systems on the determination of significance are explored, finding that while the effectiveness of the artifact systems play a role significance detections, the count of systems included in the artifact pool appears to be the most important factor; answering S-RQ4.2.

For academic researchers, the BHM-Gaussian method has the potential to provide more flexibility in making retrieval effectiveness inferences. Researchers currently propose a new retrieval model and then must evaluate it against a set of recent baselines to determine if it is better than the current state-of-the-art. In many cases these recent baselines use training data that is not available to the researcher, or the cost of implementing the baselines is prohibitive. In the BHM-Gaussian framework described, since every available run can be included in the inferential analysis as an artifact, and the system effects are moderated by the average system effect by partial pooling, the direct comparison between particular systems is less important. If the artifact pool contains systems with similar effectiveness to the current state-of-the-art baselines, then the researcher can make inferences against that run as a reference point.

4.3 Bayesian Risk On Many Systems

The experiments in Section 4.1 described the advantages of modeling paired risk-adjusted scores as Skew-Normal distributions. That was motivated by the observation in Chapter 3 that the asymmetry of the risk-adjusted scores can impact the inferential outcomes of statistical tests. Section 4.2 described how the Gaussian model could be extended to support multiple system comparisons using Bayesian hierarchical modeling, as a precondition to modeling risk-adjusted scores over multiple systems. This section combines both the BHM-Gaussian model and Paired-Skew-Normal approach explored in Section 4.1 to form BHM-Skew-Normal, allowing for risk inferences over many systems with multiple comparison correction, answering S-RQ4.3: *How does the Bayesian prior affect inferential results using risk-adjusted scores* *for one-to-many comparisons?* In defining and analyzing the performance of a Bayesian risk model over many systems, it can be compared to existing frequentist risk models, answering S-RQ4.4: *How do Bayesian and frequentist credible and confidence intervals differ when performing risk-adjusted evaluations?*

4.3.1 Methodology

How to extend the BHM-Gaussian model to support inferential comparisons of risk on many systems is now discussed. Recall that the BHM-Gaussian model takes standard IR effectiveness scores from multiple systems and models them as a Gaussian distribution. In contrast to the BHM-Gaussian model, the Paired-Skew-Normal model takes paired risk-adjusted scores relative to the champion system. Therefore, to model the risk-adjusted scores over many systems for inferential purposes, the risk transformation must be applied in terms of the observed effectiveness score, not the difference in scores relative to the champion system.

Algorithm 4.4 extends the Paired-Skew-Normal approach to handle multiple comparisons in a similar approach to the BHM-Gaussian method. The highlighted lines show the differences between the BHM-Skew-Normal and Paired-Skew-Normal, where the hierarchical terms are incorporated into the location parameter of the Skew-Normal distribution. Note that although the response distribution honors the overall shape of the risk-adjusted scores, the system α_i hierarchical term is still modelled as a Gaussian distribution. Specifying Gaussian hyper-priors for group effects in hierarchical modeling is the standard approach currently, where alternative distributions are discouraged as it is uncertain whether they may affect the validity of partial pooling.[4] With that, the risk-adjusted scores are also modeled using the BHM-Gaussian method to provide a comparison to the BHM-Skew-Normal method.

Risk Transformation. To compute risk-adjusted scores over many runs simultaneously and enable comparison against the Champion, a minor change is made to how the risk-adjustment is applied to the scores, so that the effectiveness scores modeled are in terms of their *absolute* score, not the *score difference.* This is achieved by calculating:

$$\mathit{challenger}'_j = \mathit{challenger}_j - \max\{(r-1)(\mathit{champion}_j - \mathit{challenger}_j), 0\} \tag{4.5}$$

where risk penalties are applied to the challenger's score relative to the champion's score. The $r - 1$ term is used to be consistent across the thesis, with the philosophy that when $r = 1$, then no risk penalty has been applied. To provide an example for Equation (4.5), consider a situation where losses count three times as much as gains ($r = 3$), and a champion and challenger system have the scores 0.4 and 0.1 respectively for that same topic; the challenger system is 0.3 points behind the champion and therefore incurs a penalty. The challenger 0.1 topic score is adjusted to $0.1 - ((3 - 1) \cdot (0.4 - 0.1)) = -0.5$. Alternatively, if a champion and challenger system on another topic had the scores 0.2 and 0.6 respectively, the challenger

[4] The limitation of specifying the hierarchical terms as Gaussian distributions is discussed at `https://github.com/paul-buerkner/brms/issues/231`, accessed on 17th May 2021.

Algorithm 4.4: The BHM-Skew-Normal posterior log-density accumulator. The highlighted parts describe the added components to the Paired-Skew-Normal approach to make it a hierarchical model.

```
Input: The array $S$ of IR effectiveness scores of many systems, with current chain
       (or, initial) proposals for $\langle b, \sigma, \lambda, \mu_\alpha, \alpha_i, \mu_\beta, \beta_j \rangle$ used to explore the posterior
       distribution.
Output: The accumulated log-posterior density for the given inputs, for the MCMC
        sampler to probabilistically determine whether remaining in the current
        position or moving to this location is optimal.
                                        // Accumulate prior density using brms defaults
$location \leftarrow median(S)$
$scale \leftarrow \max\{mad(S), 2.5\}$
$lp \leftarrow student\_t\_lpdf(b, 3, location, scale)$
$lp \leftarrow lp + student\_t\_lpdf(\sigma, 3, 0, scale)$
$lp \leftarrow lp + normal\_lpdf(\lambda, 0, 4)$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\alpha, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\alpha_i, 0, 1)$                  // Standard normal prior for $\alpha_i$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\beta, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\beta_j, 0, 1)$                  // Standard normal prior for $\beta_j$
                  // Parameterize the skewness in terms of mean and standard deviation
$\delta \leftarrow \lambda / \sqrt{1 + \lambda^2}$
$\omega \leftarrow \sigma / \sqrt{1 - 2/\pi \times \delta^2}$
                                                       // Accumulate likelihood density
for $y_{ij} \in S$ do
    $\hat{\alpha_i} \leftarrow \mu_\alpha \times \alpha_i$          // Model interaction of current system and average system
    $\hat{\beta_j} \leftarrow \mu_\beta \times \beta_j$             // Model interaction of current topic and average topic
    $\mu_{ij} \leftarrow b + \hat{\alpha_i} + \hat{\beta_j}$                                  // Set linear predictors
    $\xi_{ij} \leftarrow \mu_{ij} - \omega \times \delta \times \sqrt{2/\pi}$
    $lp \leftarrow lp + skew\_normal\_lpdf(y, \xi_{ij}, \omega, \lambda)$
end
return $lp$
```

retains that absolute score of 0.6 without penalty. Equation (4.5) is applied to every run before the BHM-Skew-Normal model is applied. The champion system is included in the model where it is unaffected by the risk penalty. In combining the absolute risk-adjustment approach above with hierarchical modeling, the approach is named BRisk$^-$, and is empirically evaluated in Section 4.3.2.

Experimental Setup. The experimental conditions of the multiple system risk inference scenario closely follow the risk parameters of Section 4.1, and multiple system testing scenario in Section 4.2. The same evaluation scenario involving a champion system, a set of challengers, and a pool of artifact systems, datasets, and evaluation metrics are used. For risk analysis, $r = 5$ is used for tabulated results, and $r \in \{1, 2, 5, 10\}$ is used to evaluate the sensitivity of the results to the risk parameter.

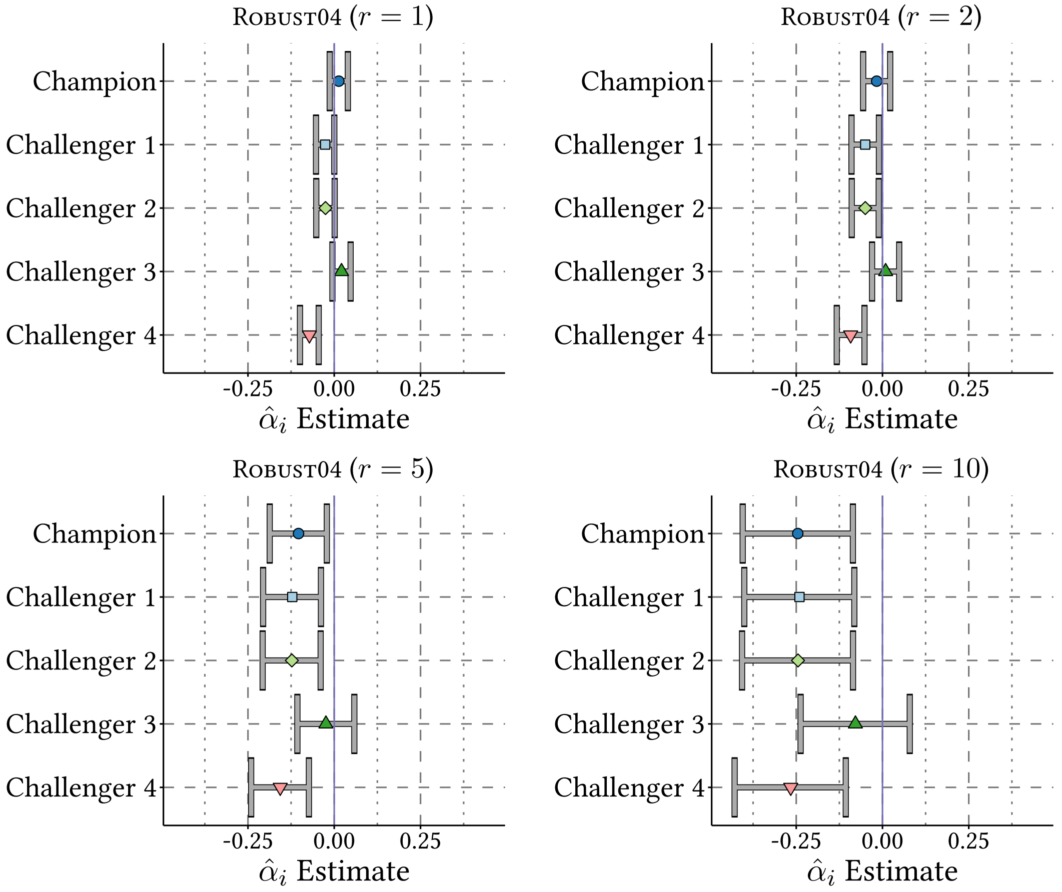


Figure 4.10: Parameter estimates of risk-adjusted scores of challengers and artifact systems compared to the champion system for various risk levels using the explored BRisk$^-$ approach.

4.3.2 Experimental Analysis

BRisk$^-$ System Inference. To answer S-RQ4.3, the system effect parameters of the BRisk$^-$ approach are compared when the risk parameter is varied on each dataset. Figure 4.10 shows the outcome of calculating BRisk$^-$ on ROBUST04 on the 50 topic subset for the risk levels that are typically explored: $r \in \{1, 2, 5, 10\}$, noting that the artifact systems are also included in this model, but their presentation is removed to focus on the champion vs. challengers scenario. Although the BRisk$^-$ approach was executed for $r \in \{1, 2, 5, 10\}$ on all three datasets of ROBUST04, TREC17, and TREC18, the results are only shown for ROBUST04 as the system orderings by risk are consistent across all datasets.

When $r = 1$, no risk is applied and the x-axis corresponds to (a reflection of) the y-axis on the top-left panel (ROBUST04 50 topics) of Figure 4.6 on page 112. At $r = 2$, Challenger 4 remains more risk-sensitive than Challenger 3 when compared to the champion system. When $r = 5$, the credible intervals begin to overlap, and they remain that way for $r = 10$. When $r = 10$, the system effect size point estimate of the champion system catches up to Chal-

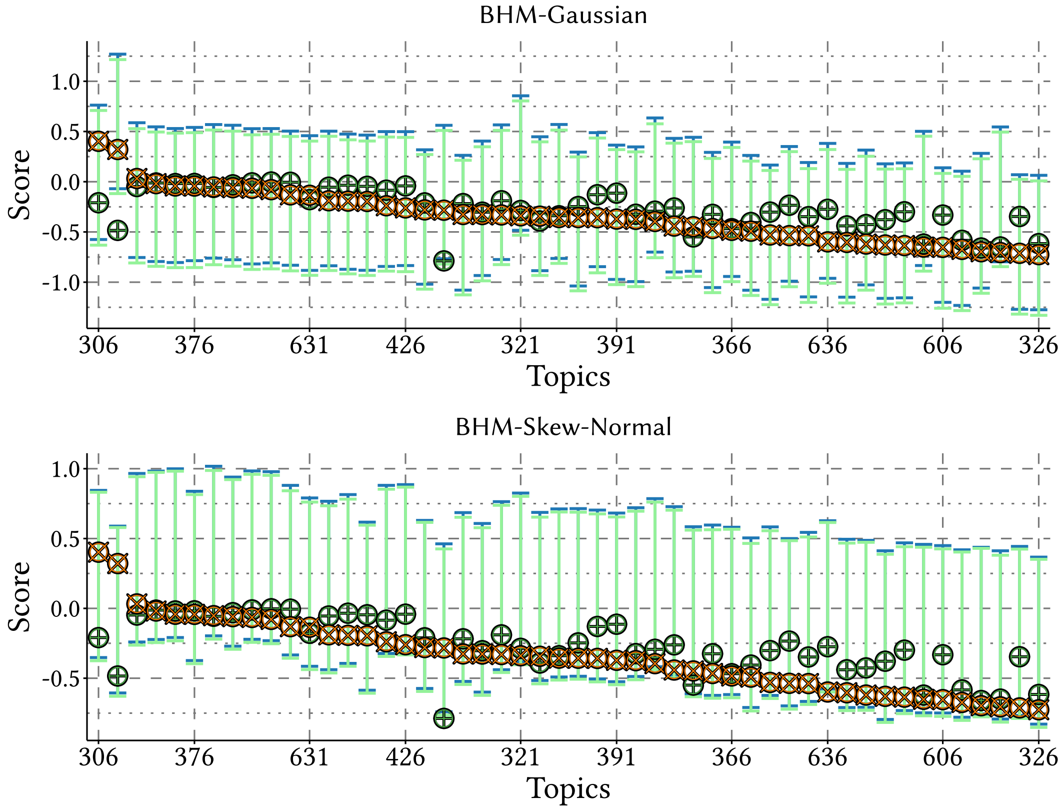


Figure 4.11: Per-topic BRisk$^-$ $r = 5$ predictive intervals over all systems for a BHM-Gaussian (in blue) against the BHM-Skew-Normal model (green bar). Challenger 3 scores are ordered from most-risky to least across topics with an orange cross, and the corresponding Champion score is shown with a green crosshair. Recall that the Champion score is unaffected by the risk penalty, and is therefore always in a state of reward (*Score* < 0).

lenger 1 and Challenger 2, while the point estimate for Challenger 4 remains marginally superior. Surprisingly, when the risk-adjusted scores were modeled using either BHM-Skew-Normal or BHM-Gaussian approaches, the outcomes were consistent across all datasets.

BRisk$^-$ Per-Topic Risk Forecasting. As Challenger 3 was consistently the riskiest choice identified in Figure 4.10, exploring where its risk-adjusted scores fall in the predictive interval over every system should thus be illuminating. Although the difference between the BHM-Gaussian and BHM-Skew-Normal approaches were not substantial when exploring the relative system effect sizes, the predictive intervals take into account the uncertainty around the overall score. Figure 4.11 compares the observed score from the champion system when compared against the risk-adjusted score for Challenger 3, using $r = 5$. Scores in the positive direction indicate risk; conversely negative scores indicate reward compared to the champion system. The topics on the horizontal axis are ordered from most risky to least risky for Challenger 3.

On the BHM-Skew-Normal plot in Figure 4.11, the risk score for the champion system is predicted to give the highest reward across each topic forecast most of the time, indicating that the model is describing the observed data well, since its scores are not penalized. On the other hand, the BHM-Gaussian plot is unable to resolve lower score values particularly well, conflating the original system scores with the risk-adjusted scores. Observed risk-adjusted scores from the Challenger 3 run are consistently higher across topics, due to it being riskier on average than the champion run. Of particular interest is the second-most risky topic on the BHM-Skew-Normal plot where the observed score of the champion system is within the predictive interval, and in the BHM-Gaussian plot, the observed score is far outside the interval. While there are topics on which Challenger 3 performs marginally better, nothing can be decidedly inferred overall that would suggest when $r = 5$ it is a better choice than the champion system.

BRisk^{-} In Context. The BRisk^{-} method is now placed in context with existing risk methods to address S-RQ4.4, as well as the paired bias-corrected method BCa^{-} proposed in Chapter 3. Table 4.5 on page 125 shows results for both paired and many system risk testing approaches for various challenger vs. champion scenarios. The analysis begins by first exploring the paired methods available: URisk^{-}, TRisk^{-}, and BCa^{-}. (Both URisk^{-} and TRisk^{-} were described in Section 2.5.) Recall that URisk^{-} is a descriptive statistic, TRisk^{-} an inferential one, and that BCa^{-} is inspired by TRisk^{-} in Chapter 3 using the bias-corrected accelerated bootstrap. Note that the TRisk^{-} overlay is not corrected for multiple comparisons. This serves as a point of comparison for multiple correction, where the appropriate reporting of a BCa^{-} value over multiple systems adjusts the confidence interval using Bonferroni correction (excluding artifacts as hypotheses to correct for). In general on different corpora, TRisk^{-} reports more significant inferences than the BCa^{-} ones, an expected outcome given BCa^{-} inferences are both Bonferroni-corrected and adjusted for bias. The outcomes of each testing method are in agreement on Challenger 3 for both Robust04 and TREC18, and on Challenger 4 for the TREC17 collection. These results are shown to contextualize the many system risk testing results.

Table 4.5 shows the outcomes of the new one-to-many risk overlay BRisk^{-} against the many-to-many overlays ZRisk^{-} and GeoRisk^{-}. The BRisk^{-} values correspond to the limits of the intervals plotted in Figure 4.10 on page 120, recalling from Section 4.3.1 that this method implicitly corrects for multiple comparisons, and that artifact systems are being used in these models. From the experiments, the BRisk^{-} method appears to be less sensitive to significant detections of risk than alternative approaches. The only significant result reported was that of Challenger 4 vs. the champion run on TREC17, which every other significance testing result registered as significant in Table 4.5 on page 125. Interestingly, systems that TRisk^{-} registered as significant such as Challengers 1 and 2 on TREC17 are not significant when using any of the bias-corrected approaches. When considering the many-to-many risk overlays, as every z-value is greater than 2.0 on the ZRisk^{-} overlay, it is the most predisposed risk measure to indicate risk for all systems for $r = 5$ (excluding artifacts). However, ZRisk^{-} scores retaining only information about the variance and shape of the risk distributions. When GeoRisk^{-} is used to include score magnitude into the ZRisk^{-} value, Challenger 4 becomes the most desirable system. But, adding the score magnitude information back into the risk measure comes at the cost that it removes the ZRisk^{-} inferential interpretation.

4.3.3 Discussion

Although the Paired-Skew-Normal approach proved useful in the one-to-one comparison explored in Section 4.1, extending it to support multiple system comparisons did not yield the same practical improvement. This is conjectured to be due to Gaussian modeling of system effects; a current constraint when multi-level modeling is concerned in Bayesian inference, providing an answer to S-RQ4.3. As can be seen in Figure 4.10, despite no risk-adjustment being applied to the champion system, the variance of the champion system estimate still increases from the contribution of the risk-adjusted score effect across all other systems, as the champion has been adjusted by partial pooling in the BHM model. As all systems in a BRisk^{-} setting except one (the champion run) have had the risk transformation applied, the average system effect distorts the champion credible interval in a way that makes drawing conclusions more difficult. Nevertheless, evaluating risk with multiple systems might still be useful for forecasting unseen scores at a per-topic observation level.

The results presented in this section saw that BRisk^{-} may be less powerful than expected. Although the BRisk^{-} point estimates match the true risk-adjusted values being modeled and are not incorrect in any way, the partial pooling approach used when modeling risk values directly impeded the ability to achieve significance compared to the alternative methods explored in Table 4.5 on page 125. The BRisk^{-} method appears unlikely to be more powerful than the alternative paired measures compared in our experiments. In the one significant outcome reported by BRisk^{-} in Table 4.5, all other methods agreed that Challenger 4 posed no risk for $r = 5$ against the champion system on TREC17.

4.4 Conclusion

This chapter has explored the unsolved problem of extending inferential IR risk methodology to support multiple system comparisons with correction for the false discovery rate. Novel Bayesian risk models were developed iteratively on IR scores; specified on both theoretical grounds and through applying the R package `brms` to the problem. The first iteration Paired-Skew-Normal was contrasted against Paired-Gaussian to model the standard one-to-one risk scenario. The more simple Paired-Gaussian model is extended in BHM-Gaussian to support modeling standard IR effectiveness scores, where the data is not skewed. Finally, the BHM-Skew-Normal model is developed to support the more complex risk analysis scenario where the data is skewed (identified to be the case in Chapter 3). To evaluate the novel Bayesian models involving a champion system with many challengers, a small-scale example was defined and then expanded upon using several newswire test collections and their corresponding set of reference (artifact) systems. The Bayesian risk testing methodology BRisk^{-} was defined and empirically evaluated against the existing risk overlay approaches. This chapter is the first work in IR on defining Bayesian models for multiple system inference, and opens the door to future evaluation work for many other goals.

Chapter 3 identified that the risk-adjusted score distribution was skewed, which may impact the validity of inferences assuming a Gaussian distribution. In the Chapter 3 results, using the bias-corrected accelerated bootstrap BCa^{-} method, the skewness of the score distribution was corrected for, but it increased the uncertainty in the inferences. In Section 4.1, the Paired-Skew-Normal model was developed to address the skewness issue and to support the one-to-one risk analysis scenario, which could later be extended to support many to many comparisons. Surprisingly, the Paired-Skew-Normal model was found to increase the certainty in the mean risk-adjusted score inferences compared to the Paired-Gaussian model. This was due to the Paired-Skew-Normal model being able to better isolate the skewness of the score distribution, leaving more certainty in the credible intervals for the location parameter; answering S-RQ4.1.

When extending the paired study to the more complex many-to-many system comparison scenario, including artifact systems enabled the model to be both sensitive to differences in effectiveness and generalizable. Since inferences in the many system scenario are informed by a set of reference systems, the outcomes of inferential analyses naturally change as the number of artifact systems increases, which was observed in the experiments in Section 4.2; answering S-RQ4.2.

In another unexpected turn of events, although modeling paired risk inferences as Paired-Skew-Normal in Section 4.1 resulted in tighter credible intervals, and the Section 4.2 inferential approach extending the paired approach to support multiple system comparisons was sensitive to differences in effectiveness, the BRisk^{-} method was less sensitive to detectable differences in risk than the alternative risk methods. Still, the BRisk^{-} multiple comparison method could provide practical benefit for forecasting risk score observations in an exploratory topic-wise data analysis, answering S-RQ4.3. When BRisk^{-} was compared against other inferential risk overlay approaches, although it was weaker in its determination of significance, for the one significant difference registered, it fully agrees with the frequentist determinations and the point estimates do correlate well with the alternative risk overlays, answering S-RQ4.4. These sub-questions answer the thesis question RQ4 in full, describing an approach that models risk-adjusted scores over multiple IR systems, with correction for the false discovery rate. Chapter 5 follows up on the sensitivity issue on Bayesian risk analyses involving multiple systems by computing risk inferences post-hoc, rather than prior to modeling the data.

	Run		One vs. one				One vs. many		Many vs. many	
	System	AP	URisk$^-$	TRisk$^-$	BCa$^-$		BRisk$^-$		ZRisk$^-$	GeoRisk$^-$
Robust04	Champion	0.274	∅	∅	∅		−0.103	[−0.187, −0.021]	12.424	−0.405
	Challenger 1	0.323	−0.024	−1.408	−0.024	[−0.067, 0.020]	−0.122	[−0.206, −0.038]	10.230	−0.433
	Challenger 2	0.322	−0.026	−1.581	−0.026	[−0.066, 0.016]	−0.123	[−0.207, −0.039]	9.311	−0.430
	Challenger 3	0.264	0.105	2.976	0.105	[0.042, 0.236]	−0.024	[−0.107, 0.058]	11.036	−0.394
	Challenger 4	0.380	−0.071	−2.345	−0.071	[−0.128, 0.031]	−0.156	[−0.241, −0.073]	10.654	−0.471
TREC17	Champion	0.210	∅	∅	∅		0.017	[−0.065, 0.099]	26.029	−0.383
	Challenger 1	0.289	−0.052	−2.077	−0.052	[−0.100, 0.031]	−0.025	[−0.107, 0.057]	23.033	−0.443
	Challenger 2	0.290	−0.053	−2.114	−0.053	[−0.100, 0.034]	−0.025	[−0.107, 0.057]	22.798	−0.443
	Challenger 3	0.216	0.015	1.817	0.015	[−0.001, 0.043]	0.029	[−0.052, 0.110]	25.969	−0.388
	Challenger 4	0.572	−0.352	−11.100	−0.352	[−0.419, −0.257]	−0.263	[−0.349, −0.179]	23.555	−0.624
TREC18	Champion	0.236	∅	∅	∅		−0.116	[−0.207, −0.024]	17.446	−0.387
	Challenger 1	0.301	−0.040	−1.882	−0.040	[−0.091, 0.014]	−0.151	[−0.244, −0.060]	16.118	−0.434
	Challenger 2	0.300	−0.042	−2.047	−0.042	[−0.092, 0.008]	−0.153	[−0.245, −0.061]	15.393	−0.432
	Challenger 3	0.231	0.065	3.468	0.065	[0.027, 0.123]	−0.058	[−0.151, 0.033]	18.925	−0.387
	Challenger 4	0.459	−0.165	−2.791	−0.165	[−0.266, 0.071]	−0.261	[−0.354, −0.169]	19.748	−0.548

Table 4.5: The AP effectiveness scores of each of the challenger and champion systems, combined with the output of each of the one vs. one risk overlays considered. The risk level is fixed at $r = 5$; figures in blue are statistically significant for the overlay in that column. BRisk$^-$ requires special attention to whether the system estimate CI of a challenger system overlaps with the champion system on its respective corpus.

5

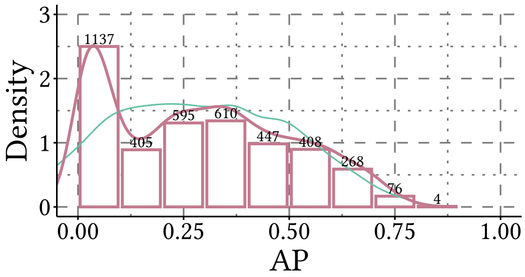
Bayesian Post-Hoc Risk Analysis On Many Systems

Evaluating whether the improvements in effectiveness outweigh the potential losses of swapping an existing ranker with an alternative model remains a foundational problem in IR. Chapter 3 explored the characteristics of risk-adjusted score distributions for two systems at a time, finding that when the losses are multiplied by a constant factor r, the asymmetry of the distribution leads to concerns about whether statistical tests assuming normality of their mean scores can reliably yield risk inferences. BRisk^{-}, described in Chapter 4, applied those findings to model risk-adjusted scores using a tailored Bayesian approach; addressing the multiple comparison problem when many systems are considered. The decision to model multiple risk-adjusted system scores as Skew-Normal instead of applying a generic nonparametric frequentist test relates to Kitchen's [113] observation that parametric tests are more powerful when their assumptions are met. However, the outcomes of BRisk^{-} in Section 4.3.2 where risk-adjusted scores were directly modeled over multiple systems had the unexpected effect that larger r values caused greater uncertainty; when interpreting the relative risk of each system against a baseline.

This chapter aims to address how to improve the sensitivity of risk-inference over multiple systems while maintaining the Bayesian hierarchical modeling multiple comparison correction. Rather than modeling risk-adjusted scores directly, the chapter explores modeling the standard IR effectiveness scores, and computing risk on the predicted values of the Bayesian model; also known as the *posterior predictive distribution* (PPD). Although the risk-adjusted score distribution is skewed, the predictive intervals are found by sorting the replicates and selecting the 2.5th and 97.5th percentiles of the PPD, with the median value representing the middle of the distribution. By virtue of that approach, risk inferences can be derived from the PPD in a nonparametric way, and the skewness can be handled implicitly. It may have been more difficult to achieve significance when modeling risk directly in BRisk^{-} as the unadjusted champion system is pooled with the other adjusted systems; subsuming the variance from the other risk-adjusted systems. That difficult to detect significance does not affect the validity of detected significance claims, but instead indicates that the test may be underpowered. If com-

	Topics				
	301	306	311	316	…
s_1	0.084	0.132	0.290	0.652	…
s_2	0.084	0.109	0.399	0.668	…
s_3	0.347	0.043	0.429	0.666	…
s_4	0.076	0.125	0.519	0.667	…
$s_{\dots}$	…	…	…	…	…

(a) AP scores from Robust04; systems ordered most effective to least by mean.



(b) A density plot of the true AP score distribution of every observation recorded left (pink line), and a random PPD draw from a Gaussian model imputed by that data (thin green line).

Figure 5.1: A graphic showing AP scores from all systems and a 50-topic subset on the Robust04 collection, where the true distribution is plotted in the density plot on the right. A Gaussian (or, "normal") Bayesian model is used with the goal of making inferential comparisons between systems. A thin sanity-check line shows a random draw from the posterior predictive distribution (PPD) of a simulated model with system and topic effects included. Draws from the PPD of a well specified model would follow the shape of the true distribution, but important details such as the bias towards low AP scores on Robust04 are lost.

puting risk inferentially over many systems as a post-hoc step on IR scores modeled directly provides a PPD to derive inferences from, then the focus must be on improving the modeling of the IR scores. An inaccurate PPD model of standard IR scores will consequentially impact the validity of risk inferences downstream.

Section 4.2 notionally applied a BHM-Gaussian model as a prior for modeling IR scores, as it is the most widely used approach in the literature for techniques such as ANOVA. Figure 5.1 presents a tabulation of all IR AP scores from Robust04 systems (those that survived the cull of the bottom 25% of systems by AP [202], that is), where the true distribution of scores is plotted in the density plot on the right. A thin line representing a draw from the PPD of a BHM-Gaussian model with system and topic effects included is presented, showing the difference between the true distribution and the PPD. For example, the original scores reflect bimodality with a bias towards lower scores, however, this PPD draw happens to simulate fewer AP values less than 0.125 than are actually observed. Figure 5.1a shows that AP scores below 0.125 are common even in the top four performing systems (by mean). The difference between the BHM-Gaussian model and the true distribution may further implicate the validity of risk inferences made on those generated values. (An alternative model for AP scores over many collections is explored later in the chapter in Figure 5.10 on page 158.) Due to these modeling concerns, this chapter explores novel BHM models for describing many IR systems simultaneously for inferential purposes, where the models can be used to generate risk inferences post-hoc.

This chapter breaks from the tradition of inferentially modeling IR effectiveness scores with pre-applied risk transformations, and instead computes risk as a summary statistic on PPD score replicates (mentioned in Section 2.3.4 on page 42). The most common use-case of a PPD is a sanity check, to determine whether simulated values from a Bayesian posterior have the same distributional properties as the original data that was modeled [27, 79, 87, 139]. Using the PPD for inference is usually unnecessary as the variables of statistical interest are already separated into regressors of the model equation of the posterior distribution. However, the PPD synthesizes the posterior distribution into replicate values of the modeled data, which enables risk to be computed on these values, and consequently, risk inferences. The typical application of the PPD used in an inferential scenario is in calculating the count ratio of an event, for example, the System A measure is greater than the System B measure, which is the Bayesian analogue of a p-value [2, 27, 132, 183] that Carterette [42] previously explored in an IR context. Whether computing risk on each draw of the PPD and interpreting the outcomes provides more utility than BRisk^{-} is unknown and forms the basis of this chapter. In particular, the results presented here reinforce the skewness experiments identified in Chapter 3, and implicitly accounts for them.

Contributions. This chapter addresses the thesis research question and sub-questions:

Research Question (RQ5): *How can standard IR scores be modeled over multiple systems (with multiple comparison correction) to improve the sensitivity of multiple system risk inference?*

S-RQ5.1: *How do increasingly sophisticated models affect the assessment of system dominance?*

S-RQ5.2: *How can Bayesian score replicates be used to infer multi-system risk, without directly modeling the adjusted scores?*

S-RQ5.3: *How does the sensitivity of posterior predictive risk compare with modeling risk-adjusted scores directly?*

As outlined by the research questions, this chapter presents the first investigation into using the predictive posterior distribution for risk inferences over multiple systems with multiple comparison correction, and explores the inferential power afforded by using more sophisticated Bayesian models of IR effectiveness scores.[1] From what can be seen in Figure 5.1 on the previous page, it is worth investigating whether the discrepancy between modeled scores and real scores in a multiple system inference scenario can be narrowed by novel fit-for-purpose models. As the complexity of the models increases, so too does the power of the system inferences (S-RQ5.1). An initial round of experimentation towards modeling RBP gain hierarchically is also presented, giving rise to substantial gains in power, but puzzling changes in system rankings relative to the more traditional inferential modeling approaches, warranting future work. Following an exploration of the behavior of different models in a BHM context, a new methodology for calculating risk inference using the posterior predictive

[1] This work appeared in R. Benham, A. Moffat, and J. S. Culpepper. Bayesian System Inference on Shallow Pools. *Proc. European Conf. on Information Retrieval (ECIR)*, pages 209–215, 2021.

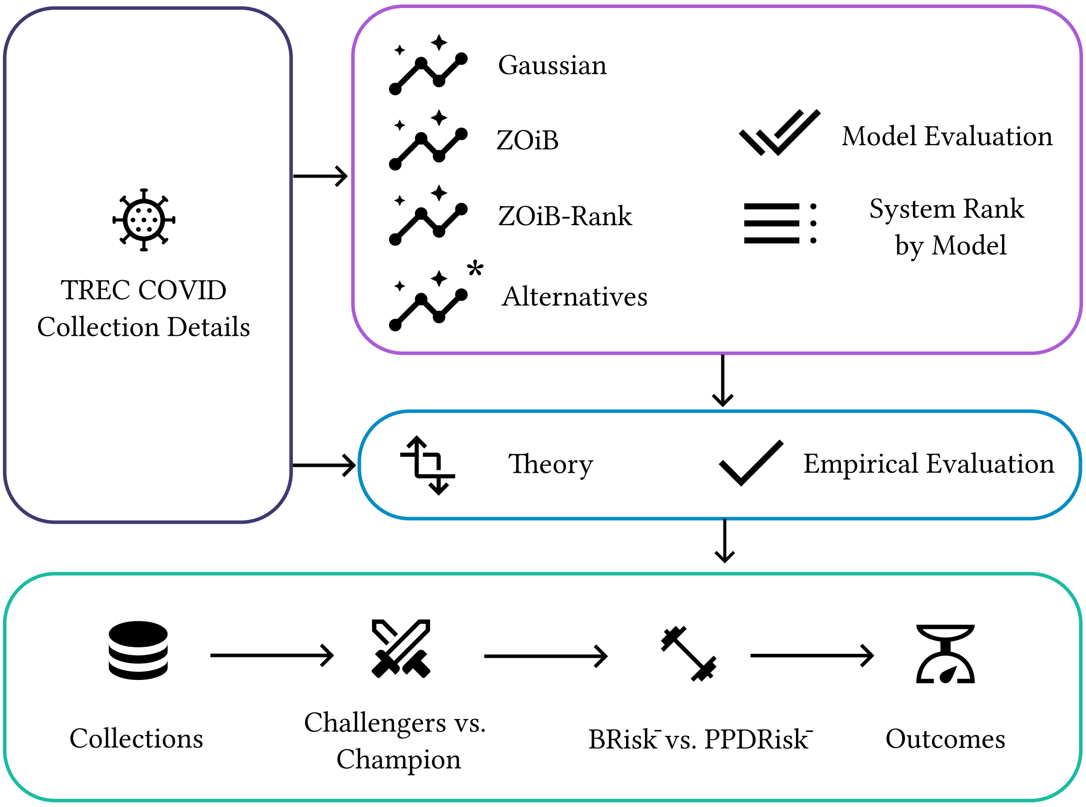


Figure 5.2: A road-map of the chapter, showing the progression from exploring system modeling to risk modeling with the posterior predictive distribution (PPD).

distribution is proposed and evaluated using the TREC COVID dataset (S-RQ5.2). It appears from the results of the experiment that the proposed approach to computing risk PPDRisk$^-$ may be able to provide more powerful inferences than that of BRisk$^-$ the approach described in Chapter 4, warranting further analysis. Revisiting the BRisk$^-$ experiments in Chapter 4, PPDRisk$^-$ appears to yield more significant results than BRisk$^-$, while retaining its benefits (S-RQ5.3).

Figure 5.2 presents a road-map of this chapter. The chapter begins by analyzing the results of the recent TREC COVID collection, where the RBP metric is used instead of AP, as Lu et al. [123] find it to be more robust in the presence of shallow judgment pools. Results of the track are presented, showing how uncertainty in score estimates still impedes analysis of system dominance without any statistical inference, where the dominant system is the one that is effective on average across all topics. In this scenario, greater computational modeling effort may be justified if risk inference is desired, as simplistic score images could further exacerbate the error between the modeled scores and the true score distribution. The sensitivity of two different Bayesian models on these effectiveness observations is interpreted and compared, introducing a novel approach using a zero-one inflated Beta distribution named BHM-ZOiB. Using the proposed methods for modeling RBP scores, Section 5.2 on page 149 describes the methodology for computing risk on posterior predictive score replicates denoted PPDRisk$^-$, with the TREC COVID dataset used as an example. Section 5.3 commencing on page 156 revisits the BRisk$^-$ experiment comparing it with the new PPDRisk$^-$ approach, and finally the chapter is concluded in Section 5.4 on page 163.

5.1 Inference On Shallow Pools

Chapter 4 explored using BHM to compare multiple systems inferentially. Retrieval effectiveness scores were modeled using generalized linear mixed models (GLMM) on AP, to address the risk asymmetry problems identified in Chapter 3. (As a reminder, GLMMs extend the gLMM methodology to link linear models to different response distributions.) The goal of this section is to understand how statistical models imbued with enhanced domain knowledge of IR effectiveness scores compare in their assessment of which ranker is the most effective. These models can be used to produce posterior predictive distributions that can be used to compute risk inferences, that might be more powerful than the BRisk$^-$ approach explored in Chapter 4. Of particular interest is the effect that collections with shallow judgment pools can have on the volatility of inferential assessments, as the recent TREC ad-hoc collection received shallow judgments in incremental stages, TREC COVID 2020. Advancements in this area may provide further utility in the event of future exercises utilizing residual collection pooling [165].

This section addresses S-RQ5.1: *How do increasingly sophisticated models affect the assessment of system dominance?* First, a recent scenario is explained where performing an evaluation campaign using a shallow judgment pool has uncertainty in relative system rankings before any statistical inference has occurred. With that done, an exploration is conducted into whether Bayesian data analysis on such collections can better capture that uncertainty. Then, the provisionally accepted set of Bayesian priors used to compare multiple IR system effectiveness is documented and validated through the application of a range of sanity checks. Conditional on that acceptance, the respective detection sensitivity for system model difference is then compared.

5.1.1 Motivation

The rationale for the research conducted in this chapter on exploring different models stems from participation in the TREC COVID 2020 IR track.[2] The evaluation campaign had five rounds (technically six rounds, when including the three runs that were used to form an initial judgment pool selected by the track organizers), where judged documents from past rounds are successively removed from the collection and new topics are added. Table 5.1 on the next page details the rapid growth in COVID-19 research throughout this assessment period. The collection of interest in this research is the first round set, in which two manual

[2] This work appeared in the workshop paper R. Benham, A. Moffat, and J. S. Culpepper. RMITB at TREC COVID 2020. In *TREC COVID Bibliography* (`https://ir.nist.gov/trec-covid/bib.html`), available on arXiv: `http://arxiv.org/abs/2011.04830`

	Round	Date	Documents	Unique Terms	Total Terms	Topics
TREC COVID	1	10th April 2020	51,078	806,018	112,451,422	30
	2	1st May 2020	59,888	863,757	117,403,028	35
	3	19th May 2020	128,493	1,158,312	155,817,513	40
	4	19th June 2020	158,275	1,275,348	185,500,005	45
	5	16th July 2020	192,510	1,284,588	200,668,989	50

Table 5.1: Statistics for the five rounds of the TREC COVID collection, where the first round was used to explore the distributional properties of Bayesian inference on shallow pools. The TREC COVID exercise was unique in that the collection grew in size as more information about the SARS-CoV-2 virus was published.

query fusion runs were submitted by the RMIT group: `RMITBM1`, where documents were reranked for their temporal relevancy, and `RMITBFuseM2`, a control run with no time-biased re-weighting.

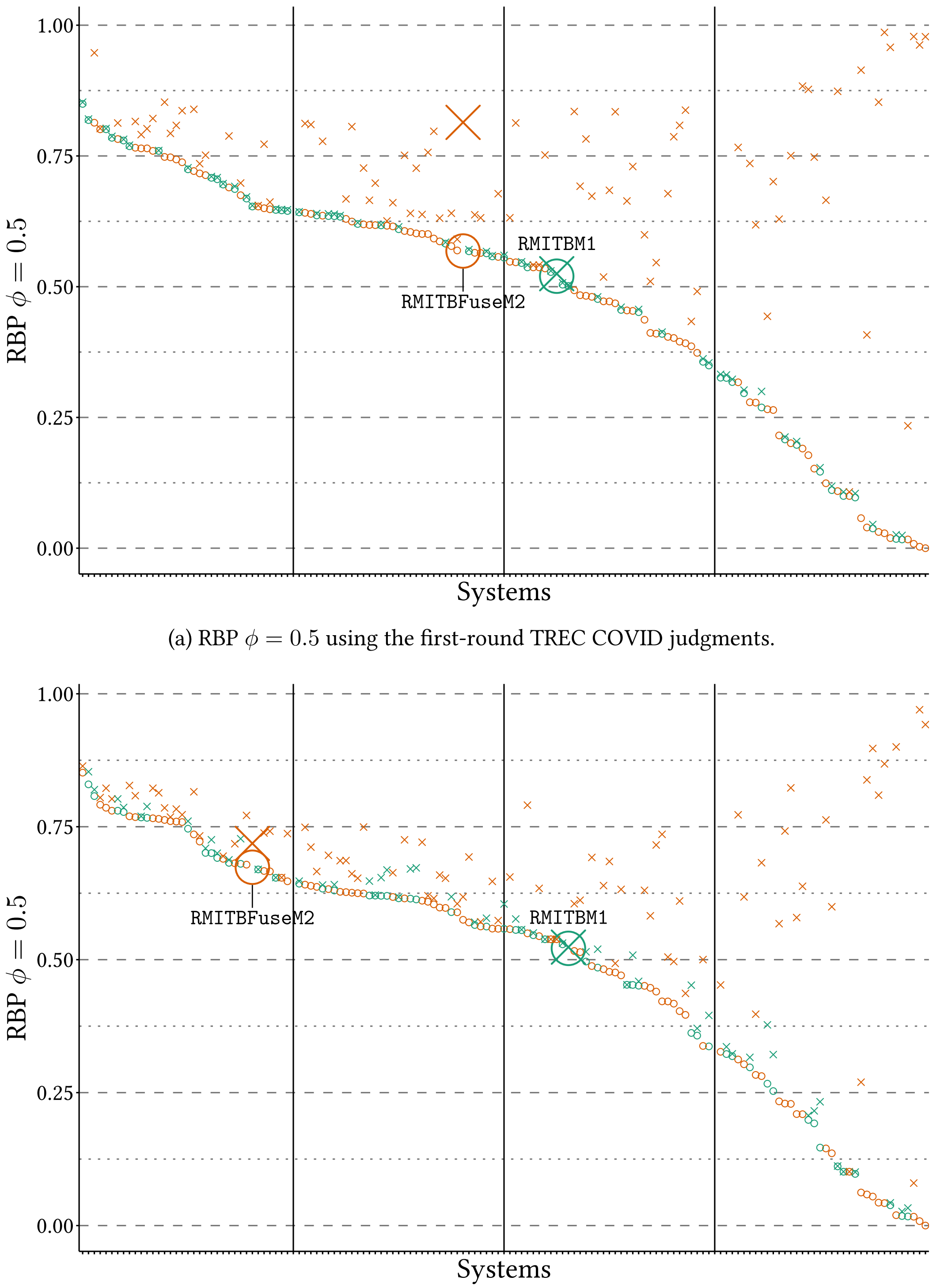
Of the RMIT runs, the former run was based on the hypothesis that the recency of COVID-19 research articles would be a more important relevance feature than traditional IR ad-hoc retrieval tasks. Ten user query variations were written for each of the 30 topics for double fusion [18]. Both runs fused the result lists of ten queries executed on eight different retrieval models implemented in Terrier v5.2, where query expansion was also applied to the queries.

Due to the high volume of participants and assessment constraints, only the first priority runs were able to be judged due to assessment budget constraints. Voorhees et al. [205, Figure 3] explain that if a system was included in the judgment pool, the top seven documents per topic were guaranteed to be judged. Jimmy Lin in the TREC COVID discussion group[3] expressed concerns upon the release of the third-round data that the shallow evaluation depth yielded many more unjudged documents than a typical IR collection[4] at rank cut-offs 5 and 10 (citing a similar evaluation exercise explored by Yang and Lin [216]) which may obscure the interpretation of evaluation metrics which treat unjudged documents as non-relevant. Since the expected evaluation depth of RBP can be evaluated using $1/(1-\phi)$ [134], when $\phi = 0.5$, the measure models the persistence of an impatient searcher with an expected search depth of just the top-two documents.

Given that scenario, Figure 5.3 on the following page shows how uncertainty in effectiveness score can impact the outcome of an IR evaluation, before any statistical inference is used. Each item along the horizontal axis corresponds to a system, and the values along the vertical axis are the mean RBP score of that system. Each system effectiveness score is reported with two values. The observable RBP score is marked with a circle, and the largest

[3] Thread: "*TREC-COVID Round 3 Analysis of Judged@k*" by Jimmy Lin in the TREC COVID Discussion Group: `https://groups.google.com/g/trec-covid/c/qWt2k8sI-pU/m/57SB8rYWBgAJ`. Note: This is a private group which requires permission to join.

[4] Data Analysis Shared Publicly on GitHub, paying special attention to the $\mathcal{J}$@10 and $\mathcal{J}$@5 columns: `https://github.com/castorini/TREC-COVID/blob/master/round3/leaderboard.csv`



(a) RBP $\phi = 0.5$ using the first-round TREC COVID judgments.

(b) RBP $\phi = 0.5$ using the complete TREC COVID judgments.

Figure 5.3: How unjudged documents in effectiveness scores can impact the perceived relative effectiveness of rankers, when evaluated against a shallow judgments and a more complete set. Measured RBP scores are marked with a circle, with the upper-bound of the score if the unjudged documents were deemed relevant is marked with a corresponding cross. These pairs are listed in green if the run was judged, and orange otherwise. Black vertical lines group each of the runs into quartiles. With more complete relevance assessments, the second-priority unpooled RMIT run is re-ranked to the top-quartile of runs.

possible RBP score is marked with a cross, the latter computed by assuming every unjudged document to be relevant. Systems marked in green were systems that received judgments over their top-7 documents [205], whereas systems marked in orange were not pooled. Looking first at the left panel, Figure 5.3a shows that the first-round judgment set has not generalized well to unpooled systems. The second priority run `RMITBFuseM2` scored slightly higher than `RMITBM1`, but the uncertainty in the score indicates that it could have been a top-quartile run. When the complete judgment set was formed it was confirmed that `RMITBFuseM2` was top-quartile:

> "The largest change in the relative ranking of runs is the `RMITBFuseM2` run which rises 33 ranks when using P@5 as the measure (21 ranks by NDCG@10, 7 ranks by AP and none for bpref)."
>
> — Voorhees [201, p. 2]

Using the RBP $\phi = 0.5$ metric, Figure 5.3b confirms that `RMITBFuseM2` is a top-quartile run, moving up 35 system rankings. These results show that comparing unjudged runs on judgment sets formed using shallow pools may yield misleading results, as Lu et al. [123] demonstrate may be problematic when exploring the relationship between evaluation and pooling depths on different kinds of metrics. If unjudged runs are excluded, it is interesting to see which systems can be statistically separated, noting that 30 topics are in use, which is a small-enough sample size to warrant the traditional testing approaches [180, 197]. Classic frequentist testing approaches assume that the system-topic scores are draws from a population distribution of system-topic scores and derive inferences based on that. But, Figure 5.3 on the previous page shows how fragile the evaluation from a shallow collection can be. If `RMITBFuseM2` can rise 35 places on a metric if more judgments were made available, then drawing inferential conclusions from synthetic scores derived from the first-round collection may be problematic.

Using Bayesian inferences, practitioners can constrain inferences based on the systems and topics measured with dependent probabilities, as well as a prior belief about the score distribution, making it an interesting application for the evaluation of IR systems based on shallow relevance judgment data. A hypothesis explored in this chapter is that Bayesian models which better honor the distributional properties of IR scores may be able to identify which system is more effective with greater statistical power. Urbano and Nagler [195] show in their experiments involving four distributions and a range of different selection criteria that there was rarely one statistical distribution best describing continuous IR effectiveness scores. However, these results are likely to be sensitive to collection, metric, and simulation methods used, and the simulation aspect is an emerging area of research. With that, choosing an acceptable model with a range of initially acceptable prior parameters and allowing the likelihood to update the prior (setting weakly-informative priors) is key to avoid the risk of skewed inferences. A typical property of IR effectiveness scores is that their range is defined as $y_{ij} \in [0, 1]$, for an effectiveness score of a metric on system i and topic j.

Urbano and Nagler [195] explored using a Beta distribution to model IR effectiveness values, among other distributions. Of interest is whether IR effectiveness values could be more flexibly represented by a modified Beta distribution instead of a Gaussian (as is traditionally done when multiple system analyses are conducted on IR scores). The Beta distribution is known to flexibly model values between zero and one:

> "The Beta distribution is very flexible for modeling data that are measured in a continuous scale on the open interval $(0, 1)$ since its density has quite different shapes depending on the values of the two parameters that index the distribution."
>
> — Ospina and Ferrari [144, p. 112]

The Zero-One Inflated Beta distribution [144] extends the Beta distribution to handle values in the $[0, 1]$ interval, which may be important in the search for a prior that accurately models the true distribution of scores, as scores of exactly zero or one do occur as a matter of routine in IR evaluation. Selecting a more informative prior than a Gaussian distribution may yield a more sensitive multiple-system inference methodology through hierarchical modeling; benefiting multiple-system risk inference methodology overall.

5.1.2 Statistical Models

Having motivated an exploration of more sophisticated statistical models combining BHM with GLMMs, the TREC COVID first round dataset is used with the RBP $\phi = 0.8$ metric, instead of $\phi = 0.5$ as in Section 5.1.1, because an evaluation depth of 7 is guaranteed for pooled runs. Using the fact that RBP scores are bound between zero and one, that information is used to establish the Bayesian priors explored in the upcoming statistical analyses. Of the 1,260 RBP $\phi = 0.8$ score observations recorded, 39 were zero (3.1%), and 17 were one (1.3%). (Although an RBP score s of one is asymptotically impossible, observations with values greater than 0.9999 are rounded to 1 for practical purposes.) To assist in potential future reproducibility exercises, while also mitigating the risk of programming error impacting the statistical analyses, the `brms` R programming language package is used as an interface to the `Stan` statistical programming language which is used to specify the models.[5] In these simulations, the default weakly-informative priors in `brms` are used when either of the Gaussian and ZOiB families are selected (these are made explicit shortly), which are auto-scaled with MCMC to be credible fits against the observed score values. The additional benefit of using the first round TREC COVID dataset is that it is a smaller dataset than a typical IR test collection, making iterative model development less computationally expensive. (The approach is further shown to support inference on larger collections on AP, shown later in Section 5.3 on page 156.) The MCMC parameters identified in Section 4.1.2 were found to be appropriate for all the models explored in this chapter; 12 chains, run at 12,000 iterations each, where the first 6,000 iterations were discarded as burn-in.

[5] Code to reproduce available at: `https://github.com/rmit-ir/bayesian-shallow`

Using the 42 pooled runs submitted to 2020 TREC COVID Track, the statistical outcomes when treating observed RBP score values are compared, assuming either Gaussian or Zero-One Inflated Beta (ZOiB) distributions. In addition to modeling RBP scores, directly modeling the RBP gain on a per-document basis (cutting each system-topic ranking to the pooling depth of 7 documents) is also explored. This is inspired by the Carterette [42] methodology where document relevance was modeled explicitly, where this approach models the RBP gain instead. These new Bayesian models will be compared against the Gaussian distribution that is typically used in IR evaluation. The BHM-Gaussian method is similar in response distribution to the gLMM approach for the many systems case [74] and the t-distribution on pairs of systems [40]. Note that it is the differences in per-topic effectiveness scores between two systems that are studentized; beyond those score pairs for multiple system comparisons, many pairs of tests are run and corrected for. Therefore this exercise cannot guarantee that one approach gives inferences that are more "truthful" than others, as such a proof does not exist. The bottom 25% of pooled systems were discarded, to avoid comparisons being performed against erroneous runs, as done by Voorhees and Buckley [202].

Linear Model. The traditionally used Gaussian distribution is a point of reference for other experimental modeling choices. This is the same model as was defined in Section 4.2 on page 104, and its use assumes that the underlying distribution of RBP values is normally distributed, as functions of a system and topic effect.

The BHM-Gaussian model has the parameterization:

$$\begin{aligned} y_{ij} &\sim N(\hat{\alpha}_i + \hat{\beta}_j, \sigma^2) & b &\sim t(3, \mathit{median}, \max\{\mathit{mad}, 2.5\}) \\ \hat{\alpha}_i &= \omega_{\alpha,\alpha_i}\mu_\alpha + (1-\omega_{\alpha,\alpha_i})\alpha_i & \mu_\alpha, \mu_\beta, \sigma &\sim t(3, 0, \max\{\mathit{mad}, 2.5\}) \\ \hat{\beta}_j &= \omega_{\beta,\beta_j}\mu_\beta + (1-\omega_{\beta,\beta_j})\beta_j & \alpha_i, \beta_j &\sim N(0,1)\,, \end{aligned} \tag{5.1}$$

where y_{ij} is an RBP effectiveness score parameterized by topic j and system i. Partial pooling is applied to the topic and system effects (β_j and α_i respectively) in $\hat{\beta}_j$ and $\hat{\alpha}_i$ [84], where the $\omega_{Y,y}$ pooling factor controls the influence of the population effect relative to observed group effect. Note, the "population" in this instance refers only to topics that were measured. The partial pooling equation takes the form

$$\omega_{Y,y} = 1 - \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_y^2}\,. \tag{5.2}$$

Gelman et al. [89] show that a frequently applied posterior validation technique is to plot the density of draws from the PPD against the original distribution and analyze discrepancies. In Figure 5.4 on the following page the original scores (thick line) are plotted against multiple independent draws from the PPD (thin lines). After using MCMC, the posterior has converged towards a distribution that describes some characteristics of the underlying effectiveness data, but is also incapable of modeling the two major peaks of the underlying score distribution when the score is high or low. That is because Gaussian values are uni-modal

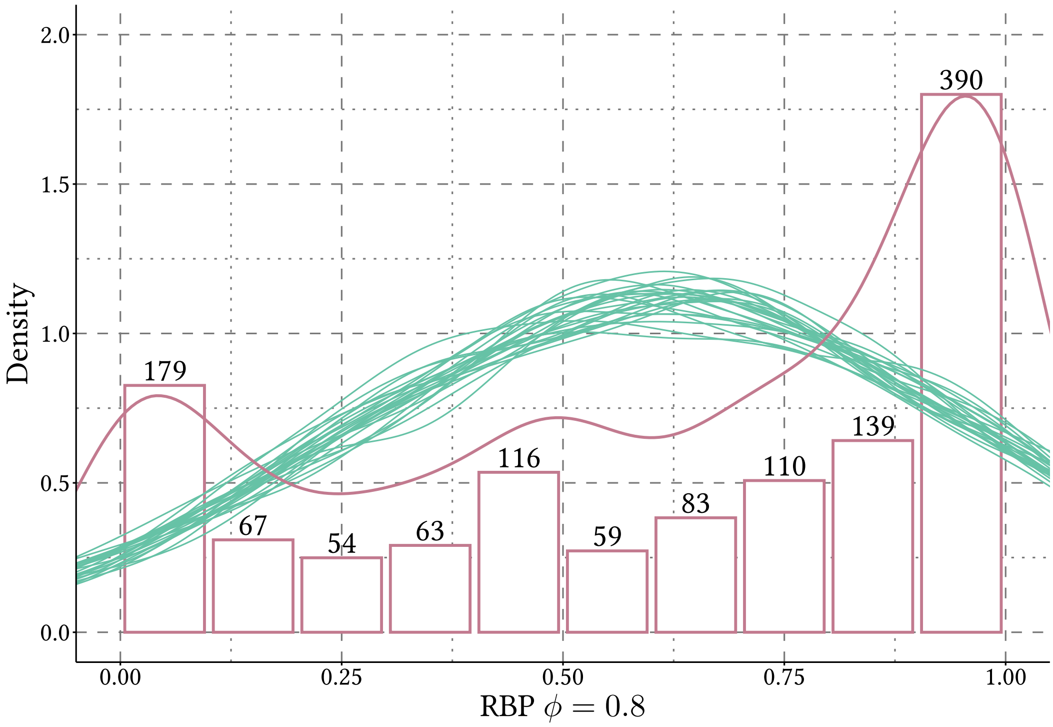


Figure 5.4: Draws from the posterior predictive distribution (thin lines) from the BHM-Gaussian statistical model against the true score distribution (pink line), where all system-topic scores are included in a model scored using RBP $\phi = 0.8$. Each RBP score observation is also displayed in equal width bins. Where the PPD draws are closest to the true distribution, this indicates good model fit.

and defined in the range $(-\infty, \infty)$. Regardless, demonstrating whether the RBP effectiveness scores are approximately Gaussian or not is less important than finding a more accurate model, and falls out of the scope of this research.

Algorithm 5.1 on the next page describes the log-posterior density subroutine used for MCMC for the BHM-Gaussian hierarchical model, where modeling decisions that specifically treat each observation as BHM-Gaussian are highlighted in green. As inputs, the subroutine takes an ordered array of effectiveness scores. As many systems are included in this array, systems are ordered lexicographically and grouped by system, where the corresponding topic scores are sorted in ascending order by topic number. Also input into the subroutine is a vector describing the current proposal state of the MCMC sampler for the parameters being modeled. (The MCMC sampler will randomly select values as initial proposals in the initial case, making this vector easy to supply.) The log-posterior accumulator variable is assigned on line 3, which is the returned result used by the particular MCMC kernel employed to determine whether the current proposal set is an acceptable representation of the posterior distribution. The subroutine is grouped into two parts, where the density of the weakly-informative priors are accumulated between lines 3 and 8, and the hierarchically modeled system and topic effects are computed in the likelihood part in lines 8 and 14. A linear pass is made over each of the original observations, where the pooled estimate of the current system

Algorithm 5.1: The BHM-Gaussian posterior log-density accumulator. The highlighted parts describe the components of the hierarchical model that make it specifically model a Gaussian random variable.

```
Input: The array $S$ of IR effectiveness scores of many systems, with current chain
       (or, initial) proposals for $\langle b, \sigma, \mu_\alpha, \alpha_i, \mu_\beta, \beta_j \rangle$ used to explore the posterior
       distribution.
Output: The accumulated log-posterior density for the given inputs, for the MCMC
        sampler to probabilistically determine whether remaining in the current
        position or moving to this location is optimal.
                                        // Accumulate prior density using brms defaults
$location \leftarrow median(S)$
$scale \leftarrow \max\{mad(S), 2.5\}$
$lp \leftarrow student\_t\_lpdf(b, 3, location, scale)$
$lp \leftarrow lp + student\_t\_lpdf(\sigma, 3, 0, scale)$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\alpha, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\alpha_i, 0, 1)$               // Standard normal prior for $\alpha_i$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\beta, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\beta_j, 0, 1)$                // Standard normal prior for $\beta_j$
                                              // Accumulate likelihood density
for $y_{ij} \in S$ do
    $\hat{\alpha}_i \leftarrow \mu_\alpha \times \alpha_i$        // Model interaction of current system and average system
    $\hat{\beta}_j \leftarrow \mu_\beta \times \beta_j$            // Model interaction of current topic and average topic
    $\mu_{ij} \leftarrow b + \hat{\alpha}_i + \hat{\beta}_j$                              // Set linear predictors
    $lp \leftarrow lp + normal\_lpdf(y, \mu_{ij}, \sigma)$               // Probability density function
end
return $lp$
```

and topic is computed, and combined with the error (or intercept) term b to form a linear combination of μ for the GLMM. That μ is then used to calculate the relative likelihood that the observation y_{ij} is generated with σ.

ZOiB Model. To address the above point where the statistical model may benefit from including more precise details on the original score distribution, a Beta distribution can be used to model a rate in the range $(0, 1)$, and a ZOiB distribution extends that range to $[0, 1]$. A Beta distribution has two shape parameters α and β, enabling it to be more expressive for values between 0 and 1 than a Gaussian distribution.[6] Therefore, RBP scores are modeled with the

[6] `https://rdrr.io/cran/brms/man/brmsfamily.html`, accessed on 8th June 2021.

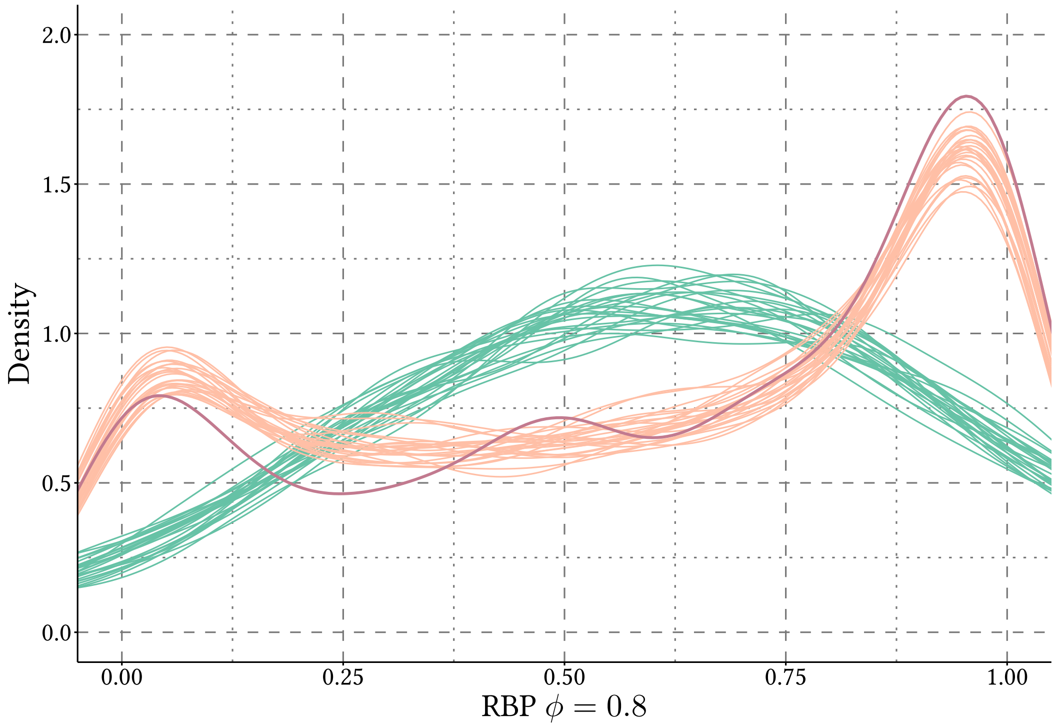


Figure 5.5: Draws from the posterior predictive distribution (thin lines) from the BHM-ZOiB statistical model against the original score distribution (pink line), where all system-topic scores are included in a model scored using RBP $\phi = 0.8$. Each RBP score observation is also displayed in equal width bins. Where the PPD draws are closest to the original distribution, this indicates good model fit. BHM-Gaussian draws from Figure 5.4 on page 137 are also displayed for comparison.

BHM-ZOiB parameters:

$$y_{ij} \sim \begin{cases} \pi_0 & \text{if } y_{ij} = 0 \\ (1-\pi_0)(1-\pi_1)\beta(\mu_{ij} \cdot \phi, (1-\mu_{ij})\phi) & \text{if } 0 < y_{ij} < 1 \\ \pi_1 & \text{if } y_{ij} = 1 \end{cases}$$

$$\begin{aligned} \text{logit}\, \mu_{ij} &= \hat{\alpha}_i + \hat{\beta}_j & \mu_\alpha, \mu_\beta &\sim t(3, 0, \max\{\mathit{mad}, 2.5\}) \\ b &\sim t(3, \mathit{median}, \max\{\mathit{mad}, 2.5\}) & \pi_0, \pi_1 &\sim \beta(1,1) \\ \hat{\alpha}_i &= \omega_{\alpha,\alpha_i}\mu_\alpha + (1-\omega_{\alpha,\alpha_i})\alpha_i & \phi &\sim \gamma(0.01, 0.01) \\ \hat{\beta}_j &= \omega_{\beta,\beta_j}\mu_\beta + (1-\omega_{\beta,\beta_j})\beta_j & \alpha_i, \beta_j &\sim N(0,1)\,, \end{aligned} \tag{5.3}$$

where ϕ is the precision parameter of the Beta distribution β to be modeled with a Gamma distribution (another `brms` default), π_0 and π_1 are the Bernoulli probabilities that a score will be respectively exactly zero or exactly one, and μ_{ij} is logit transformed to link the linear parameterization (described in BHM-Gaussian) to the Beta distribution.

The corresponding posterior predictive check for the BHM-ZOiB on RBP $\phi = 0.8$ scores is presented in Figure 5.5. The PPD score replicates are visually more representative of the multi-modal components of the original distribution and appear to fit better than the BHM-

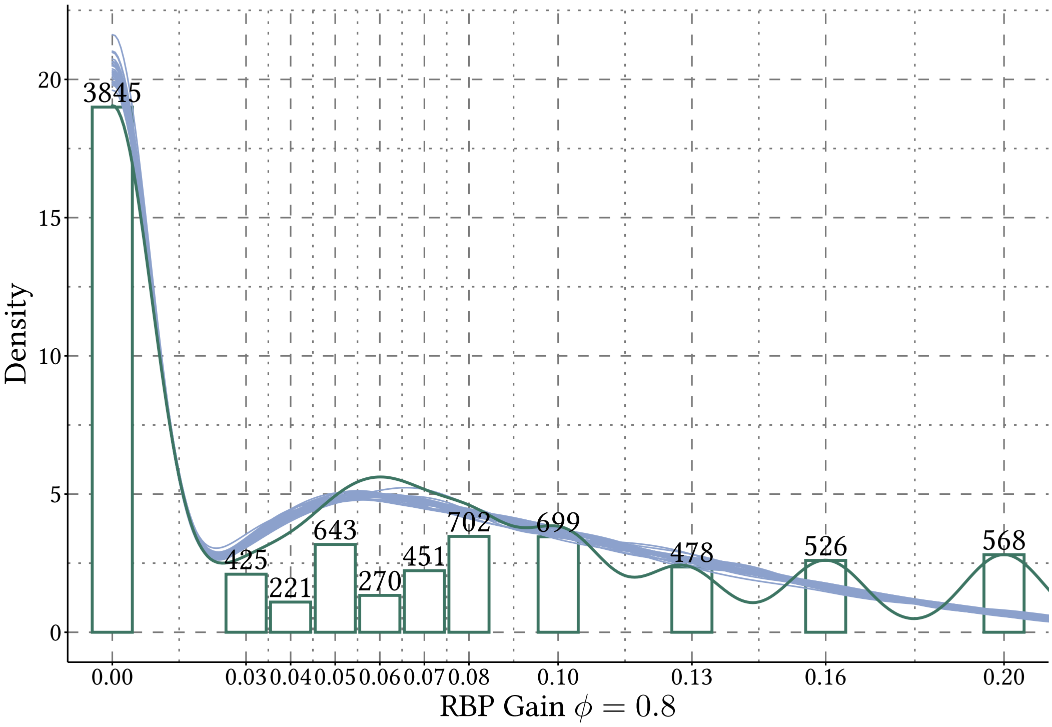


Figure 5.6: Density plot of RBP $\phi = 0.8$ document gain values (individual addends from rank positions in the RBP sum) along with posterior predictive draws from a BHM-ZOiB-Rank model, for every pooled system, and every topic. Beyond including rank as a predictor, all other variables are as already noted in connection with Figures fig. 5.4 on page 137 and fig. 5.5 on the previous page.

Gaussian method shown in Figure 5.4 on page 137, but it remains to be seen how important these details are in the context of statistical inferences. (The interested reader is invited to peek ahead at Figure 5.10 on page 158 where the same comparison is made between models on the official TREC AP measure on many collections, and observe Table 5.2 on page 144 for quantitative determinations of which model fits better.) As the next model iterates on BHM-ZOiB, both algorithms are explained below in Algorithm 5.2 on page 142.

ZOiB-Rank Model. Carterette [42] showed that one of the benefits of using more sophisticated models to perform system inference was the ability to model system dominance at a document-level in a system-topic ranking. Since each of the systems were fully judged to depth seven in the experiments involving the TREC COVID dataset [205], the BHM-ZOiB model can be extended to model y_{ijk} per-position RBP gain scores by including k as a rank parameter, modeled as a population effect. Although RBP gain scores are different to RBP scores distributionally, the values are bound to the range zero and one and the Beta distribution is quite expressive. BHM-ZOiB-Rank is therefore a small modification: logit $\mu_{ijk} \sim N(\hat{\alpha}_i + \hat{\beta}_j + k, \sigma_y^2)$. (Carterette [42] used the similar Quasi-Binomial distribution to model RBP gain scores, a response family that is not available in brms.)

The posterior predictive check of modeling RBP gain values at a system-topic-rank level is plotted in Figure 5.6 on the previous page. The up-bends in Figure 5.6 on the previous page are places where each relevant document-rank combination arises as a power of ϕ, for example, a relevant document found at the first rank would yield a score of $0.2 \times 0.8^0 = 0.2$, and in the second position $0.2 \times 0.8^1 = 0.16$, 0.128, and so on. The replicated scores capture an overall approximation of of the shape of the original score distribution, allowing BHM-ZOiB-Rank to be provisionally accepted for inferential purposes. In critiquing the BHM-ZOiB-Rank model, the distribution of RBP gain scores is evidently multi-modal, where approximately half of the density of the highest gain values for RBP $\phi = 0.8$ are not captured by the model. Whether there is value in treating RBP gain scores as discrete and identifying a better fitting probability density function is discussed after the system inferences have been explored against the BHM-Gaussian and BHM-ZOiB models.

Algorithm 5.2 on the next page shows the process used for computing BHM-ZOiB or BHM-ZOiB-Rank by highlighting the salient details in a spot-the-difference scenario compared to Algorithm 5.1 on page 138. Note that the green highlighted lines that are specific to modeling the response distribution as Gaussian are not present in Algorithm 5.2 on the next page, but the same unhighlighted code remains. As BHM-ZOiB is a proper subset of the steps required to compute BHM-ZOiB-Rank, the latter requires inclusion of all highlighted code, whereas BHM-ZOiB exists in its own right with only the orange highlighted lines included. In observing just the BHM-ZOiB lines and contrasting against Algorithm 5.1 on page 138, the relevant priors defined in the above model equations are in commensurate places in the prior / likelihood parts of the subroutine. A key difference is in the likelihood computation in the for loop, where a link function (logit) is used to enable the linear combination of IR effects to be mapped towards a non-Gaussian distribution; in this case, the zero-one inflated Beta distribution. As there are more parameters to model as inputs into the algorithm, it is clear that BHM-ZOiB comes at the trade-off that it is computationally more expensive than BHM-Gaussian.

When extending the BHM-ZOiB method to model RBP gain scores, they are input as Y, with a one-to-one R array which maps Y values to their respective rank positions. For example, an RBP gain of $y_{ijk} = 0.20$ is a document judged as highly-relevant at the top of a ranking, therefore $r_{ijk} = 1$, or the fifth ranked document may have been non-relevant; yielding $y_{ijk} = 0.00$ and $r_{ijk} = 5$. Included as an input is the $\overline{R}$ value, which is used to zero-center the k predictor to improve sampler efficiency, as k is treated as fixed rather than a random effect. (It would be if the goal were to ascertain whether a particular rank is significantly different in influencing RBP gain, but the goal in this research is to explore system dominance, and the extra computation to model the term hierarchically is not worthwhile in this case). As can be seen on line 14, the current rank value is zero-centered and multiplied by the current state of the k value supplied by the sampler, continuing with the standard behavior of BHM-ZOiB from thereafter.

Algorithm 5.2: The BHM-ZOiB (or BHM-ZOiB-Rank, if highlighted is included) posterior log-density accumulator. That is, BHM-ZOiB exists in its own right as a subset without the blue highlighting, and BHM-ZOiB-Rank is the superset of all highlighting.

Input: The array Y of IR effectiveness scores of many systems S, (or, the utility metric gain scores for each system-topic-rank combination, $G(S, k)$ where $k \in R$, the set of evaluated ranks for gain, and $\overline{R}$ to zero-center for improved sampling speed). Additionally, current chain (or, initial) proposals are supplied for exploring the posterior distribution: $\langle b, \phi, \pi_0, \pi_1, \mu_\alpha, \alpha_i, \mu_\beta, \beta_j, k\rangle$

Output: The accumulated log-posterior density for the given inputs, for the MCMC sampler to probabilistically determine whether remaining in the current position or moving to this location is optimal.

```
                                              // Accumulate prior density using brms defaults
$location \leftarrow median(Y)$
$scale \leftarrow \max\{mad(Y), 2.5\}$
$lp \leftarrow student\_t\_lpdf(b, 3, location, scale)$
$lp \leftarrow lp + gamma\_lpdf(\phi, 0.01, 0.01)$
$lp \leftarrow lp + beta\_lpdf(\pi_0, 1, 1)$
$lp \leftarrow lp + beta\_lpdf(\pi_1, 1, 1)$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\alpha, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\alpha_i, 0, 1)$                 // Standard normal prior for $\alpha_i$
$lp \leftarrow lp + student\_t\_lpdf(\mu_\beta, 3, 0.0, scale)$
$lp \leftarrow lp + normal\_lpdf(\beta_j, 0, 1)$                  // Standard normal prior for $\beta_j$
                                                                  // Accumulate likelihood density
for $y_{ij}$ (or, $y_{ijk}$) $\in Y$ do
    $\hat{\alpha_i} \leftarrow \mu_\alpha \times \alpha_i$     // Model interaction of current system and average system
    $\hat{\beta_j} \leftarrow \mu_\beta \times \beta_j$         // Model interaction of current topic and average topic
    $\mu_{ij} \leftarrow b + ([R_{ijk} - \overline{R}] * k) + \hat{\alpha_i} + \hat{\beta_j}$    // Set linear predictors
    $\mu_{ij} \leftarrow inv\_logit(\mu_{ij})$                  // Transform link function
    $lp \leftarrow lp + zoib\_lpdf(y, \mu_{ij}, \phi, \pi_0, \pi_1)$     // Probability density function
end
return $lp$
```

Empirical Model Selection. The `brms` package provides a convenient method for exploring the validity of the theoretically justified BHM-Gaussian and BHM-ZOiB models and their applicability towards modeling IR effectiveness scores with system and topic effects. As Bayesian simulation enables specifying a virtually limitless number of models, it would be remiss to suggest that the BHM-ZOiB model is the only possible choice compared to BHM-Gaussian for inferential purposes. In an attempt to find other possible candidates, focusing only on standard IR scores, the `brms` package has many different statistical distributions to choose from, with their associated weakly-informative priors for Bayesian inference vetted by expert statisticians.[7]

[7] `https://rdrr.io/cran/brms/man/brmsfamily.html`, accessed on 17th September 2022.

As the original reference model BHM-Gaussian models IR effectiveness values as if they belong to a continuous distribution, the models explored here also focus on continuous distributions. It is important to recognize that IR effectiveness scores are technically discrete variables, the common practice in IR research of treating them as continuous for inferential purposes has several advantages. The first is that it allows for the use of a wider range of metrics, as continuous distributions are more flexible than discrete ones. The second is that continuous distributions have had more attention from the statistical community, and thus, have more reliable prior specifications for Bayesian inference. Finally, the ability to preempt the running time of the MCMC sampler is a significant advantage of continuous distributions compared to bespoke discrete ones for each effectiveness metric, as the latter might have intractable convergence characteristics for differences in metric properties often perceived as minor. For example, in P@10 vs. P@100, there are 90 more discrete values to model, and that does not imply a linear increase in running time. What discrete distributions can offer in a Bayesian multiple system context is worthy of further investigation in future work.

As IR has a rich history of testing different retrieval models and assessing their viability, the reader may wish to apply the same reasoning to inferential model evaluation. The `loo` package provides a convenient way to compare the performance of different models, using the widely used WAIC measure of model fit. Taking an unbiased empirical approach towards establishing which of the supported `brms` distributions is the best fit for the IR effectiveness data, the following process was followed:

1. For each of the supported `brms` distributions, a model was specified and attempted to be fitted against the IR effectiveness data.

2. If the model was accepted for MCMC sampling, the MCMC diagnostics were inspected to ensure that the sampler produced a valid result.

3. If the MCMC diagnostics were valid for the accepted model, it was evaluated using the WAIC measure.

4. The model with the lowest WAIC value has the best fit.

The results of this empirical model selection process are shown in Table 5.2, including the mean wall time for the MCMC sampler to complete on each chain for each model. Recall that 12 chains were run for each model in 12 parallel processes with 12,000 iterations, and the mean wall time is the average of the 12 chains. Many of the distributions were rejected outright by `brms` due to the type of the observed score variable. Three distributions were accepted for MCMC sampling but then produced sampling errors, and were therefore ignored as modelling candidates. Of the considered models, the BHM-ZOiB model has the best fit for the TREC COVID RBP effectiveness data, whereas BHM-Gaussian model has the third best fit.

The BHM-Skew-Normal model explored in Chapter 4 has the second best fit. Whether it is a reasonable belief to presume skewness will be systematically present across any matrix of system-topic scores for all test collections is debatable. However, as discussed in Chapter 4,

		Sampling Errors		
brms Family	Mean Chain Time (min)	$\hat{R} > 1.0$	D. Iter	WAIC (LOO)
zero_one_inflated_beta	4.17	×	×	−656.1
skew_normal	3.28	×	×	168.9
gaussian	0.78	×	×	538.1
student	0.90	×	×	545.7
hurdle_gamma	2.25	×	×	1,583.8
hurdle_lognormal	1.41	×	×	2,458.4
gen_extreme_value	63.80	1.05	653	×
asym_laplace	28.60	×	2	×
exgaussian	1.81	×	12,833	×

The following models were rejected from sampling: poisson, negbinomial, geometric, bernoulli, binomial, beta_binomial, categorical, multinomial, cumulative, cratio, sratio, acat, Gamma, weibull, exponential, lognormal, frechet, inverse.gaussian, cox, beta, dirichlet, logistic_normal, shifted_lognormal, wiener, hurdle_poisson, hurdle_negbinomial, zero_inflated_poisson, zero_inflated_negbinomial, zero_inflated_beta_binomial, zero_inflated_beta.

Table 5.2: A comparison of many supported brms distributions, with the BHM-ZOiB model being the best fit as indicated by the WAIC measure. Distributions that were accepted for MCMC sampling but then produced sampling errors were removed as contenders for measuring model fit, and further, families that brms rejected altogether based on the type of the observed score variable is also tabulated.

when the shape parameter λ of the distribution is equal to zero then it is equivalent to the Gaussian distribution. Therefore, inferences derived from the BHM-Skew-Normal method are, in the worst case, at least as methodologically sound as using the standard BHM-Gaussian approach. With that, if BHM-ZOiB is considered to be too expensive to simulate, the BHM-Skew-Normal approach may be considered as an appropriate alternative.

The behavior of the student distribution is similar to the BHM-Gaussian model, but with a marginally greater computational cost and worse model fit according to the WAIC measure. The hurdle_gamma and hurdle_lognormal distributions converged but have poor model fit in contrast to the other explored models. No situations could be conceived where the hurdle-based models could produce more value than IR scores modeled as Gaussian given the data and evaluation metric explored.

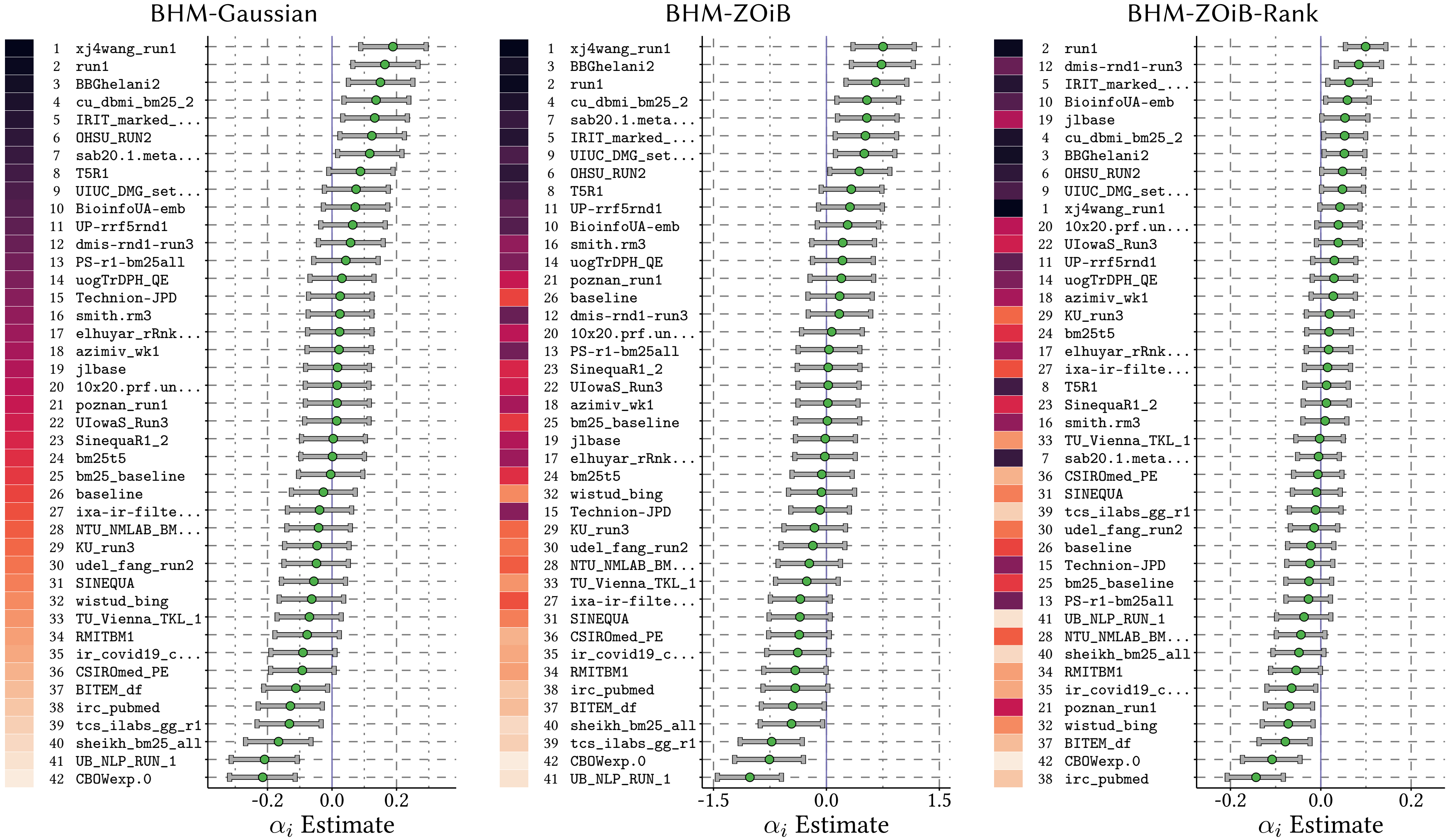


Figure 5.7: Analyzing RBP $\phi = 0.8$ system effects for each model with 95% credible intervals. Color-gradient representations of the left-most ranking accompany the numbers to the left of each run. The rank-biased overlap ($\phi = 0.9$) [210] against BHM-Gaussian is 0.877 and 0.538 respectively.

	BHM-Gaussian	BHM-ZOiB	BHM-ZOiB-Rank
Most Effective System	17	17	20
Least Effective System	23	29	31

Table 5.3: The number of times each model was able to achieve statistical significance for the most and least effective system, against any other system.

5.1.3 Experimental Analysis

Figure 5.7 on the previous page provides a head-to-head comparison of the discriminative capacity of each of the BHM-Gaussian, BHM-ZOiB, and BHM-ZOiB-Rank models, where the system effect for 95% credible intervals are listed in order of most effective to least effective system. The number to the left of each system name corresponds to the ranking that the BHM-Gaussian model gave for that particular system, based on the median parameter value within the credible interval. One drawback of using Bayesian inference is that the traditional frequentist power analysis techniques are not applicable; and using simulated data to compare between models is intractably expensive, and introduces bias from the method used to simulate the data. (This is explained in more detail shortly.) To derive a sketch of the relative power of each model, counting the number of times the best system can be distinguished from other systems, and the opposite for the worst system, may be considered an initial heuristic.

Table 5.3 shows the number of times each model is capable of distinguishing between the best and worst systems. Through inspecting how each of the credible intervals overlap, the best system can be distinguished from the 17 least effective systems, and the worst system from 23 of the most effective systems with the BHM-Gaussian model. Using the BHM-ZOiB model, the relative system orderings are permuted slightly, with the largest change being the `Technion-JPD` run being ranked 15 by the BHM-Gaussian model, but is ranked 27 using the BHM-ZOiB model and is statistically significantly less effective than the `xj4wang_run1` run according to this model. One other point of difference between the BHM-ZOiB model and the BHM-Gaussian model is that the credible intervals of the BHM-ZOiB varies more within the group of systems, as observed by the `Technion-JPD` run adjacent to the `wistud_bing` run. This may reflect how the shape of the predicted IR effectiveness values changes the credibility of the effect estimates more flexibly.

When evaluating the BHM-Gaussian model in contrast with the RBP Gain-based BHM-ZOiB-Rank model in Table 5.3 using the same procedure of counting significant differences relative to the best and worst systems, the best system can be differentiated against 20 other less effective systems, and the worst system is statistically significantly worse than 31 other systems. This heuristic approach for interpreting statistical power indicates a positive outcome for modeling at the system-topic-rank level, but the changes in system ordering against the BHM-Gaussian baseline ordering are concerning. Despite the checks and balances employed to validate the posterior of the BHM-ZOiB-Rank method, it is inconclusive whether the model is specified well and not capturing the important details of the underlying data (a type III error [197]), or whether the model is reflective of the data and too much emphasis is being placed on the relative order of the systems by their point median BHM-Gaussian estimate. For BHM-ZOiB-Rank, the 12th best system from the BHM-Gaussian model (dmis-rnd1-run3) moved up to 2nd place with BHM-ZOiB-Rank, and the run xj4wang_run1 moved from 1st to 10th based on the median estimate. Even with those substantial changes in order, these credible intervals are still overlapping on the BHM-Gaussian model. And, although the run jlbase moved from 19th place to 5th place, systems in the top-10 for BHM-ZOiB-Rank have a precision overlap of 0.8 for BHM-Gaussian, and 0.7 for the BHM-ZOiB model.

Given that BHM-ZOiB fits the score distribution better than the BHM-Gaussian counterpart and leads to more consistent system orderings than the BHM-ZOiB-Rank approach, the BHM-ZOiB model provides a balance of improved accuracy in the description of system ranking dominance compared to the BHM-Gaussian approach, and a conservative system ordering compared to the BHM-ZOiB-Rank on the first round TREC COVID dataset. Using the similarity metric RBO $\phi = 0.9$, the BHM-ZOiB model is 0.877 indicating high similarity against the BHM-Gaussian method, and 0.538 for the BHM-ZOiB-Rank against BHM-Gaussian, signalling a moderate rank correlation.

5.1.4 Discussion

Several sophisticated GLMM models have been used in combination with Bayesian hierarchical modeling to model IR effectiveness scores to: ascertain whether statistical power can be improved in a multiple comparison setting; to explore how the relative system rankings are affected by the selected statistical model; and to understand how system-topic-rank RBP gain observations can be used to perform system inference. Later in the chapter, the BHM-ZOiB findings are shown to generalize to the different metric AP, on other datasets.

Recap. The first round of the TREC COVID track is targeted as an interesting application domain, as shallow judgments are cheaper to collect and insights may be more readily transferable to other collections. Carterette and Smucker [43] found that more topics with fewer judgments was twice as efficient on assessment budgets than judging deeper into result lists for each topic, for the same statistical power. With Bayesian inference, it was previously unexplored how the paradigm stacks up power-wise against the traditional frequentist approach, where deriving enough power over fewer shallowly judged topics is the more conventional goal. As it stands, for a test collection to meet minimally acceptable statistical power requirements to produce trustworthy inferential results, 50 topics tends to be the community norm [34]. Judging documents is a laborious task, so reducing assessment costs is an important concern for all practitioners. Sakai [157] shows that hundreds or thousands of topics may be required to statistically separate a respective score difference of 0.05 or 0.02 for AP scores using an ANOVA with 10 systems for ad-hoc search on a newswire collection. Figure 5.7 on page 145 showed that including rank at the RBP gain level for BHM-ZOiB-Rank as a predictor potentially resulted in tighter intervals and hence greater statistical confidence, but further investigation is required to determine whether this was specific to the combination of systems and topics explored, or is generalizable. While none of the models considered can distinguish an RBP $\phi = 0.8$ score difference of, for example, 0.125 between `xj4wang_run1` and `T5R1`, the intervals are narrower on the BHM-ZOiB-Rank method in general. Inference based on document-level gain is an interesting avenue of future work, which may eventuate to getting more power out of less data (this is discussed further in Section 6.2.1 on page 167).

On Power Analysis. Attempting to run a more traditional power analysis on bespoke Bayesian models such as BHM-ZOiB is a challenging proposition. As the ZOiB distribution is not comparable to the Gaussian distribution, the analytical equations [160, 209] used for BHM-Gaussian models to explore power do not apply. Simulation of IR scores to explore the power of tests is a more recent approach to power analysis, but there are still some conflicting results, and Bayesian models are more expensive to compute than classic frequentist tests. Further compounding the tractability issue, Table 5.2 showed that the BHM-ZOiB model was more than five times slower than the BHM-Gaussian model to simulate.

When simulation is involved, Parapar et al. [145] contest that the t-test should not be the preferred choice, instead maintaining that consistency between collection splitting methods [58, 169, 196, 202, 224] is not a surrogate for knowing the null hypothesis, and hence that experiments which compare the agreement of tests to any other test may be biased. Urbano et al. [197] challenged the correctness of the Parapar et al. [145] approach of treating new runs for the same topics as being analogous to controlling the null hypothesis, as IR tests infer system dominance from a population of topics; and instead proposed simulations of *new topics* over the same set of systems. Urbano et al. [197] used a Urbano and Nagler [195] stochastic simulation to dynamically select the mathematical distributions that best fit the score data and minimize over-fitting, and infer a copula with respect to that marginal distribution to form new topic scores. Their results disagree with the observations of Parapar et al. [145], with the Wilcoxon and sign tests found to be the least powerful; that is, agreeing with older work on statistical power for IR.

Attempts have been made to formulate a more principled approach to power analysis of Bayesian models, such as *Bayesian Generalized Power* proposed by Kruschke and Liddell [117, Figure 12], where the key idea is to use a PPD of a Bayesian model to generate synthetic values from the original data and run subsequent tests on the predicted values to tally the ratio of tests passed. The above discussion shows that the simulation technique can bias the outcomes of the power analysis, potentially requiring a diversified set of simulation techniques to be used. All the while, each run of BHM-ZOiB may take several minutes to hours to complete depending on the size of the data and computational resources available, making simulation impractical.

Outcomes. S-RQ5.1 is answered in the affirmative, in so far as more sophisticated models do appear to allow for greater statistical sensitivity, and do not appear to yield errant relative system rankings. The BHM-ZOiB method was found to have worked well for the collection and the RBP $\phi = 0.8$ metric used here as a test case. Upcoming experiments in the chapter will show its applicability to other collections and metrics.

Suggested use-cases for BHM-ZOiB are:

- Statistically inferring whether a system outperforms another, where these experiments show that the BHM-ZOiB method even with the most pessimistic view is at least as powerful as BHM-Gaussian. Further experimentation may reveal it to be more conclusively powerful than BHM-Gaussian.
- In cases where it is important to rank / leaderboard systems, the system effects in the BHM-ZOiB method are a more accurate reflection of the underlying system ranking than taking the mean score across topics, because it honors the skewness of the underlying effectiveness scores. It is well established that skewed distributions affect the mean, and using the median to counter this fact is not well eschewed as it substantially reduces statistical power. Even so, these rankings should always be taken with a grain of salt in accordance with the uncertainty of the system effect estimates (shown in Figure 5.7 on page 145).
- A good option for creating simulated effectiveness scores that are comparable with the real ones and pre-applying multiple comparison correction for many system comparisons is the BHM-ZOiB PPD method (examined next).

Indeed, Urbano and Nagler [195] show that a one-size-fits-all model is rarely the best option from a set of statistical distributions that map well to effectiveness data. The key takeaway of this section is that there are many ways to model effectiveness data, and that there may be better ways to model the data than the BHM-Gaussian model that is commonly used in IR for multiple system comparisons.

5.2 Posterior Predictive Risk

Improving the sensitivity towards detecting significant differences in risk over multiple systems with multiple comparison correction is a key goal of this thesis. Medical applications are an example of a domain in which it is important to understand risk properties. Table 5.4 on the next page extracts a portion of the risk significance testing procedures measured in Table 4.5 on page 125 of Chapter 4 for the Robust04 collection. The proposed BRisk^{-} model was not able to achieve significance on any system for the Robust04 test collection, whereas the TRisk^{-} and BCa^{-} models were able to for the Challenger 3 vs. Champion scenario. For a BRisk^{-} system effect credible interval to be significant, it must be non-overlapping with another system. However, the credible intervals for the BRisk^{-} model in Table 5.4 all overlap. Recall from the Chapter 5 introduction the hypothesis that improved modeling of standard IR scores may lead to a more powerful method for assessing the significance of differences in risk using the PPD, rather than modeling adjusted scores directly.

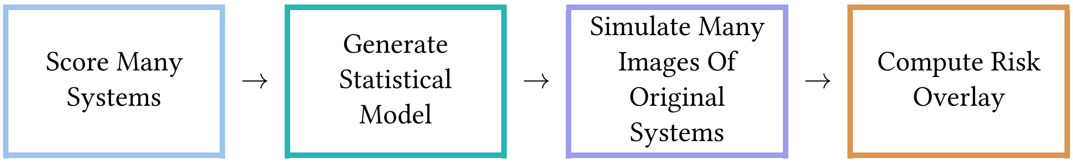
	System	AP	TRisk^{-}	BCa^{-}	(One vs. one)	BRisk^{-}	(One vs. many)
Robust04	Champion	0.274	$\varnothing$	$\varnothing$		-0.103	$[-0.187, -0.021]$
	Challenger 1	0.323	-1.408	-0.024	$[-0.067, \ 0.020]$	-0.122	$[-0.206, -0.038]$
	Challenger 2	0.322	-1.581	-0.026	$[-0.066, \ 0.016]$	-0.123	$[-0.207, -0.039]$
	Challenger 3	0.264	2.976	0.105	$[\ 0.042, \ 0.236]$	-0.024	$[-0.107, \ 0.058]$
	Challenger 4	0.380	-2.345	-0.071	$[-0.128, \ 0.031]$	-0.156	$[-0.241, -0.073]$

Table 5.4: A re-creation of the risk significance testing procedures observed in Table 4.5 on page 125 of Chapter 4, where the primary concern is whether the tests indicate a statistically significant difference in risk for $r = 5$ against the champion system with a 95% confidence/-credibility level. TRisk^{-} and BCa^{-} are paired testing methods with Bonferroni correction ($m = 5$) applied respectively, whereas BRisk^{-} model corrects for all submitted risk-adjusted TREC system scores through Bayesian partial pooling. BRisk^{-} does not achieve significance.

This section addresses S-RQ5.2: *How can Bayesian score replicates be used to infer multi-system risk, without directly modeling the adjusted scores?* Recall that BRisk^{-} defined in Section 4.3 on page 117 directly modelled the adjusted risk scores and lacked sensitivity to differences in risk compared to other methods (shown in Table 5.4). First, other related work exploiting the PPD in an IR context is discussed, and then the methodology for how the PPD can help to compute risk inferences over multiple systems is explored in Section 5.2.1 on page 152. Then, the experimental conditions of BRisk^{-} are revisited with a newly proposed methodology with preliminary results in Section 5.2.2 commencing on page 154.

Related Work. There is little related work that describes the posterior predictive distribution in an IR context, largely because the interesting quantities have already been directly modeled in the posterior distribution (the simulated set of parameters modeled, for example, the set of credible topic or system effects). However, when additional flexibility is required, the PPD can facilitate more advanced analyses and aid in the interpretation of models by synthesizing simulated values back into their original units. One such use of the PPD is related to Sakai [161], where the effect size of different systems is quantified using Glass' delta using *expected a posteriori* (EAP) values in one-to-one system comparisons. As Glass' delta is the mean difference between a baseline and an experiment divided by the standard deviation of the baseline, the interplay between the multiple comparison correction imparted by partial pooling and Glass' delta is interesting, noting that it is unclear how it should be computed in relation to the global system parameter for shrinkage.[8] The implications of computing Glass' delta with and without shrinkage applied is an interesting avenue of future work.

[8] A related unanswered question on StackExchange: `https://stats.stackexchange.com/questions/163526/effect-size-for-contrasts-in-hierarchical-bayesian-anova`


Score Many Systems
→
Generate Statistical Model
→
Simulate Many Images Of Original Systems
→
Compute Risk Overlay

System	1	2	3	4	5	6	7	…	30	Mean
xj4wang_run1	0.939	0.998	0.984	0.320	0.836	0.992	0.902	…	0.625	0.849
T5R1	0.303	0.995	0.008	0.750	0.039	0.855	0.692	…	0.999	0.724
bm25_baseline	0.906	0.744	0.946	0.500	0.143	0.746	0.994	…	0.957	0.610
CBOWexp.0	0.766	0.001	0.000	0.000	0.500	0.923	0.558	…	0.114	0.349

Chain	Iter.	Draw	$\hat{\alpha}_{xj4w}$	$\hat{\alpha}_{T5R1}$	$\hat{\alpha}_{bm25}$	$\hat{\alpha}_{CBOW}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	b	σ
9	3,663	51,663	0.21	0.11	0.01	−0.18	0.05	−0.12	−0.38	−0.27	0.64	0.29
10	3,870	57,870	0.17	0.17	−0.05	−0.22	0.22	−0.02	−0.30	−0.16	0.59	0.29
1	2,986	2,986	0.19	0.11	0.11	−0.19	0.16	−0.01	−0.28	−0.09	0.58	0.30

```
xj4w_1_51663 <- {set.seed(12345); rnorm(1,0.21+0.05+0.64,0.29)} # =1.07
```

System	Draw 51,663				Draw 57,870				Draw 2,986			
	1	2	3	4	1	2	3	4	1	2	3	4
xj4w...	1.07	0.94	0.44	0.45	1.16	0.21	0.64	0.52	0.84	0.48	0.46	1.23
T5R1	0.54	0.53	0.70	0.57	1.21	1.16	0.27	0.15	0.37	1.22	0.27	0.79
bm25...	1.29	1.00	0.34	0.52	0.67	0.04	0.75	0.39	1.19	−0.03	0.09	0.88
CBOW...	0.68	−0.04	−0.08	0.75	0.61	0.45	−0.12	0.29	0.76	0.63	0.75	−0.40

Draw	System	$\Delta 1_{bm25}$	$\Delta 2_{bm25}$	$\Delta 3_{bm25}$	$\Delta 4_{bm25}$	$\Sigma\Delta^{+}$	$5 \cdot \Sigma\Delta^{-}$	PPDRisk^{-}
51,663	xj4wang_run1	−0.22	−0.06	0.10	−0.07	0.10	−1.75	0.41
	T5R1	−0.75	−0.47	0.36	0.05	0.41	−6.10	1.42
	CBOWexp.0	−0.61	−1.04	−0.42	0.23	0.23	−10.35	2.53
57,870	xj4wang_run1	0.49	0.17	−0.11	0.13	0.79	−0.55	−0.06
	T5R1	0.54	1.12	−0.48	−0.24	1.66	−3.60	0.49
	CBOWexp.0	−0.06	0.41	−0.87	−0.10	0.41	−5.15	1.19
2,986	xj4wang_run1	−0.35	0.51	0.37	0.35	1.23	−1.75	0.13
	T5R1	−0.82	1.25	0.18	−0.09	1.43	−4.55	0.78
	CBOWexp.0	−0.43	0.66	0.66	−1.28	1.32	−8.55	1.81

Figure 5.8: An example of how the PPD can be used to compute risk distributionally. Through generating a Bayesian statistical model from the effectiveness scores of systems/topics, images of the original scores can be simulated using drawn model parameters from posterior distribution (randomly selected subset for illustration). For each draw from the PPD, a URisk$^-$ score can be computed ($r = 5$ explored), yielding the PPDRisk$^-$ distributional estimate.

5.2.1 Methodology

Until now, methods of exploring the risk characteristics in an inferential way have centered on transforming score differences into symmetric Gaussian-shaped distributions. Section 5.1.2 on page 135 described a method to simulate batches of system-topic scores for many systems at a time with multiple comparison correction applied, yielding a novel mechanism of instrumenting the risk characteristics of multiple systems inferentially. (Noting that this technique, and any other correction approach is not flawless [83, 88].) If a posterior predictive distribution is judged to be a reasonable and accurate interpretation of the underlying scores, then it can be used to interpret the spread of summary statistic values [82]. That is, for each draw from the posterior $\theta_i \sim p(\theta \mid \mathit{data})$, the set of point parameter estimates from that draw θ_i is used to form *a posteriori* replicate scores supplied to URisk: $\mathit{data}'_i \sim p(\mathit{data} \mid \theta_i)$ [118].

Figure 5.8 on the previous page describes the steps involved in generating the posterior predictive score replicates above using IR data and the URisk$^-$ equation to compute PPDRisk$^-$. To simplify explaining PPDRisk$^-$ a BHM-Gaussian model is used, however, the results are reported with the more accurate BHM-ZOiB.

Score Many Systems. The top-most table in Figure 5.8 shows a selection of TREC COVID systems with their RBP $\phi = 0.8$ effectiveness scores for a sample of the 30 topics, ordered from most effective mean score to least. The `xj4wang_run1` submitted by Wang et al. [208] that achieved the highest first round RBP $\phi = 0.8$ score used a continuous active learning model with a human-in-the-loop for training, and is included in the example to represent a top-tier system. Another effective run on RBP is `T5R1` submitted by Zhang et al. [221], that is both fully automatic and multi-stage. An initial stage uses BM25, which is then re-ranked using a Text-To-Text Transfer Transformer (T5) [149] trained on the MS MARCO passage dataset. A `bm25_baseline` also contributed to the judgment pool in the first round and is indicative of average performance of the submitted pooled systems (the point estimate of the relative system effect compared to other systems is exceptionally close to zero (the average effect) on both the BHM-Gaussian and BHM-ZOiB models shown in Figure 5.7 on page 145).[9] Finally, the least-effective system to make the judgment pool on RBP $\phi = 0.8$ was `CBOWexp.0`, which employed query expansion.[10] Focusing on the first four topics, most systems provide highly effective rankings on the first topic and volatile effectiveness on the third.

Generate Statistical Model. The sub-table (green border) in Figure 5.8 describes a sample of three random (valid, non burn-in) draws from the 72,000 MCMC simulated draws over 12 chains in the posterior distribution. System and topic effects are represented as linear combinations, with their weights denoted as α and β respectively, identified by subscript. To allow the example to fit on one page, the results of the first four topics are shown ($\hat{\beta}_{1..4}$), and the

[9] `bm25_baseline` run report: `https://ir.nist.gov/trec-covid/archive/round1/bm25_baseline.pdf`, accessed on 27th May 2022.

[10] `CBOWexp.0` run report: `https://ir.nist.gov/trec-covid/archive/round1/CBOWexp.0.pdf`, accessed on 27th May 2022.

remaining topics are omitted (but were included in the simulation of the model). A draw is identified by its chain and its current iteration in the sampler by a multiplicative mapping of each value ($9 \times 3{,}663 = 51{,}663$), where after burn-in and validation, their weights can be interpreted as part of a homogeneous posterior distribution. The next logical grouping of unitless system weights α is indicative of the system's influence on RBP effectiveness scores. The larger the weight the more effective the system is in contrast to others in the group. Weights correlate to the mean system performance (matching the ordering in the light blue sub-table), where values close to zero characterize the mean system effectiveness of the systems modeled. The topic β weights behave in the same group-wise way, where the relative topic difficulty mentioned previously is captured by the model, as $\beta_1 \gg \beta_2 \gg \beta_3$. The overall statistical error is captured in *b*, where σ is the estimate of the population standard deviation.

Note that the three random draws from the posterior distribution are only shown by way of example, and all available draws from the posterior are used when performing inferential analysis. The likelihood of getting a bad (or unlucky) draw here from this sampling is a combination of whether the draw from the posterior is outside of the credible interval (the typical researcher-defined 95% range), and whether the model which has been used to generate the draws is actually reflective of the original data. This is why it is critical for this method to be used with a well-specified model. The MCMC process acts as a filter to ensure that the posterior is populated with representative parameters.

Simulate Many Images Of Original Systems. Between the green and violet sub-tables in Figure 5.8, an R expression is presented as an example of computing a posterior predictive replicate for the first topic of the `xj4wang_run1` run, using the model weights for posterior draw $51{,}663$ listed in the table above. The seed is set to 12,345 for the Mersenne-Twister [129] uniform pseudo-random number generator which is then transformed into Gaussian space.[11] Since partial pooling has already been applied during MCMC, the $\hat{\alpha}_{\mathit{xj4w}}$ and $\hat{\beta}_{\mathit{1}}$ values have already been corrected for multiple comparison inferences, and the posterior predictive replicate represents the modeled values in the original units. As the RBP score was modeled as a linear combination of system and topic effects, the mean to supply to the `rnorm` function is

$$\mu_{51663,\mathit{xj4w},\mathit{1}} = \hat{\alpha}_{51663,\mathit{xj4w}} + \hat{\beta}_{51663,\mathit{1}} + b\,, \tag{5.4}$$

where *b* is the intercept term encapsulating statistical error. For this draw, the standard deviation σ of the population distribution is estimated to be 0.29, this is also supplied to `rnorm`. The resulting image of draw $51{,}663$ for the top system on the first topic is 1.07, which is within expectations, as the Gaussian distribution does not explicitly model scores within $[0, 1]$. Therefore, the risk of a value falling outside of this region is the combination of having an unlucky draw in combination with the effect of the pseudo-random number generator. When this process is repeated over many thousands of draws on the posterior distribution, values outside

[11] R Random Number Generator manual: `https://stat.ethz.ch/R-manual/R-devel/library/base/html/Random.html` accessed on 12th May 2022.

of this region are unusual. One benefit of using more bespoke models such as BHM-ZOiB is that posterior predictive replicates are forced to honor the $[0, 1]$ characteristic of IR scores, where the density of generated values are more in line with the original score distribution.

Compute Risk Overlay. The bronze sub-table of Figure 5.8 shows how PPDRisk$^-$ is calculated (for example purposes, noting this is a small part of the overall computation), with r set to 5 using values from the first four topics, and again using the three random draws considered above. For the purpose of risk comparison, the `bm25_baseline` is set as the champion system, and all other systems are challengers. For each draw, the image of each challenger-topic combination is subtracted from the champion-topic replicate value. For example, on draw $57{,}870$ for the first topic, run `xj4wang_run1` has the score replicate 1.16, and the champion `bm25_baseline` run has the score 0.67, making $\Delta 1_{bm25} = 1.16 - 0.67 = 0.49$. As 0.49 is positive, the challenger outperforms the champion on this topic for this draw and no risk adjustment applies. But, for the third topic the opposite is true, as $\Delta 3_{bm25} = 0.64 - 0.75 = -0.11$, and so this score difference belongs to the set of losses to be penalized by r.

When the value in the $\Sigma\Delta^+$ column is added to the $5{\cdot}\Sigma\Delta^-$ value from the same draw all divided by the number of topics, this produces a distributional estimate of PPDRisk$^-$. For example, for draw $51{,}663$ on `xj4wang_run1` against `bm25_baseline`, $-[(0.10 + -1.65)/4] = 0.41$, which is one estimate of the $72{,}000$ draws from the posterior predictive distribution. The other PPDRisk$^-$ point estimates computed in the table for illustrative purposes for are `xj4wang_run1` versus `bm25_baseline` are -0.06 and 0.13. Once all draws have an associated point estimate computed, all PPDRisk$^-$ point values are sorted, the median value is used as a representative point estimate, and the Q2.5 and Q97.5 quartiles are used to form predictive intervals for inferential purposes.

This subsection has described how to compute PPDRisk$^-$ values using a small worked example. This methodology is next used on the first round of the TREC COVID dataset to explore its application to the systems already studied above, and will later be used to revisit the BRisk$^-$ experiments in Section 5.3.

5.2.2 Experimental Analysis

The properties of the fully evaluated PPDRisk$^-$ distributions over the $72{,}000$ draws of the PPD are now explored (the product of 12 chains run with $6{,}000$ viable posterior samples after filtering out the burn-in iterations), where all pooled systems are compared against the `bm25_baseline` run for two risk parameters, $r \in 1, 2$. Recall from Chapter 4 that $144{,}000$ iterations of MCMC were performed in total to ensure an effective sample size of $10{,}000$ draws from the posterior distribution, to accurately calculate 95% credible intervals for the model parameters. To streamline the analysis pertaining directly to risk-adjusted scores and the associated PPD distribution, the ZOiB model is selected as an example of one statistical model used to generate PPD replicates. (A head-to-head evaluation of the PPDRisk$^-$ inferences supplied by the BHM-Gaussian vs. BHM-ZOiB is presented later in the chapter). After

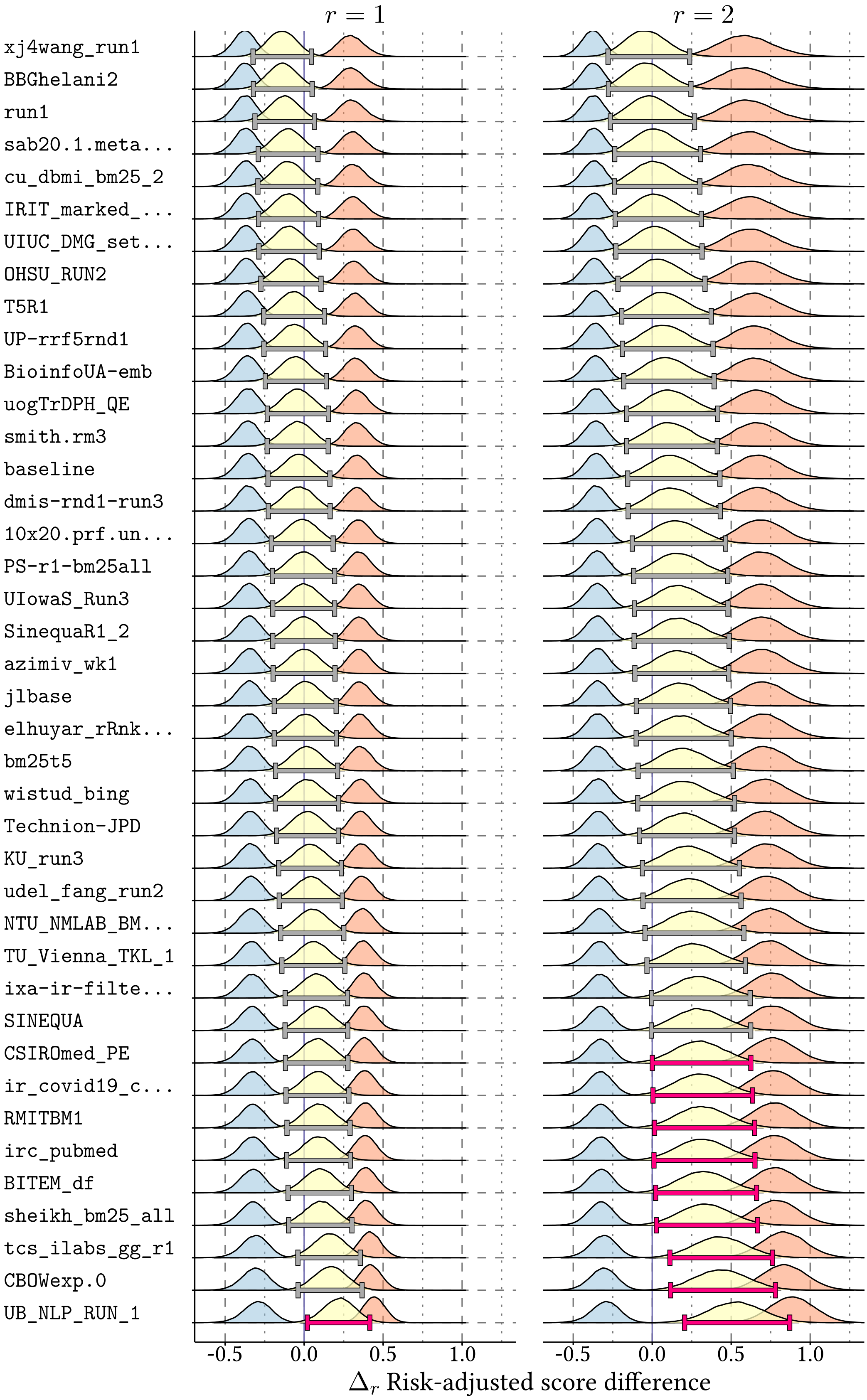


Figure 5.9: PPDRisk$^-$ density plots along with wins and risk-scaled loss distributions for $r = 1$ (left) and $r = 2$ (right). The champion run is removed as its score difference is zero.

simulating images and computing PPDRisk^{-} values for each draw, Figure 5.9 on the previous page presents the distributions of risk-free ($r = 1$) PPDRisk^{-} values against risk-sensitive ($r = 2$) evaluation. Three distributions are displayed in each r panel from left-to-right: the distribution of score differences considered a gain against the baseline system; the PPDRisk^{-} values in yellow (with an associated 95% predictive interval); and finally the losses score difference distribution is presented in red. Observe in the $r = 2$ panel, the losses distribution has been doubled in width relative to the $r = 1$ versions, to graphically aid in understanding why the PPDRisk^{-} distribution also begins to skew.

In the left-column describing a situation where no loss adjustments have been applied, `UB_NLP_RUN_1` is able to be discriminated from the `bm25_baseline` run as the predictive interval excludes zero. This is consistent with the credible interval inference plot in Figure 5.7 on page 145, in which the `UB_NLP_RUN_1` is the last item listed (per median system-effect value) in the middle panel for α_i, where the credibility interval is non-overlapping to the left of the `bm25_baseline` interval. In the right-hand column of Figure 5.9, more challengers are statistically separable as a consequence of applying the risk transformation, while still benefiting from the partial pooling applied from the hierarchical modeling employed during MCMC sampling.

5.2.3 Discussion

This section has explored a methodology to address S-RQ5.2, via including a small worked example, and is then validated using TREC COVID round 1 data. The PPDRisk^{-} approach shows promise as a more powerful methodology to determine statistically significant differences in system effectiveness compared to BRisk^{-}, as nine significant differences were detected on $r = 2$. In contrast, recall that as r increases for BRisk^{-}, the uncertainty paradoxically increases that makes it more difficult to distinguish systems.

The next section revisits the experiments used to explore the relative merit of the BRisk^{-} method against existing risk techniques. Now, the aim is to validate the rationale for using BHM-ZOiB on other effectiveness metrics, such as AP employed in the previous study in Section 4.3 on page 117.

5.3 Revisiting Multi-System Risk

The section above presented a novel methodology for computing risk inferences over multiple systems with correction, while showing promise that it may be more powerful than the BRisk^{-} method presented in Chapter 4. With that, this section addresses S-RQ5.3: *How does the sensitivity of posterior predictive risk compare with modeling risk-adjusted scores directly?* In addition to exploring the relative power of BRisk^{-} versus the newly proposed PPDRisk^{-} method, this section provides further evidence of the applicability of modeling AP, another IR effectiveness metric, towards the BHM-ZOiB model on three document collections when compared to the typical BHM-Gaussian approach.

Collection	Citation	Documents	Unique Terms	Total Terms	Topics
Robust04	[199]	528,155	664,603	253,367,449	250
TREC17	[13]	1,855,658	2,970,013	1,285,653,766	50
TREC18	[3]	595,037	1,478,198	481,432,022	50

Table 5.5: A repeat presentation of Table 4.3 on page 110 to detail the collection statistics of the experiment explored in Chapter 4, to be explored in the context of PPDRisk^{-}.

5.3.1 Experimental Setup

Section 4.3 on page 117 introduced the idea of using Bayesian inference to compute risk evaluation over many systems. By applying a risk transformation prior to generating a Bayesian statistical model using a Skew-Normal distribution, this yielded a risk inference technique denoted BRisk^{-}. To explore how PPDRisk^{-} compares in discriminative ability against BRisk^{-}, and the alternative frequentist-based approaches, an identical experimental setup is used to compare the BRisk^{-} method.

In Section 4.2.3, a champion system was selected compared to four other challenger runs of varied effectiveness. The champion was an untuned BM25 run, with Challenger 1 and 2 query expansion runs using Bo1 from Terrier 5.2, Challenger 3 was a bag of words run with sequential dependencies, and Challenger 4 was a run intentionally designed to be highly effective, taking the CombSUM fusion of the top-three submitted runs for Robust04, and reciprocal rank fusion of the top-three for TREC17 and TREC18. See Table 5.5 for statistical information about each of the test collections, as well as Table 5.6 on the next page for more details in regard to the effectiveness of each run.

Posterior Predictive Checks. Figures 5.4 to 5.6 earlier in this chapter provided visual evidence of the improved model fit of the BHM-ZOiB model compared against the BHM-Gaussian method on RBP $\phi = 0.8$ effectiveness values on the first round of the TREC COVID test collection. To explore how repeatable this outcome is on other collections and evaluation metrics, Figure 5.10 on the next page demonstrates a head-to-head comparison of the synthetic draws from the PPD of the BHM-Gaussian model against the original scores for three further collections, in contrast with the BHM-ZOiB approach on the bottom row. Included in each graph is the relative model fit WAIC score for each model, and the time taken to simulate each chain. Recall that lower WAIC scores indicate better model fit.

In the left column, the original AP scores on the Robust04 collection of runs are bimodal, with low effectiveness scores common and a second peak about the 0.33 AP score region. The bimodality of the AP scores in the Robust04 column compared to other two test collections in Figure 5.10 on the next page is unsurprising, as Robust04 is a curation of topics that have been known to be difficult for systems to identify relevant documents for in previous TREC ad-hoc exercises. For the BHM-Gaussian model, the mode with the highest density is poorly represented by the model, whereas the BHM-ZOiB alternative models both peaks.

		AP Score		
System	Description	Robust04	TREC17	TREC18
Champion	BM25	0.274	0.210	0.236
Challenger 1	DPH + Bo1	0.323	0.289	0.301
Challenger 2	DPH + DFR + SD +Bo1	0.322	0.290	0.300
Challenger 3	DPH + DFR + SD	0.264	0.216	0.231
Challenger 4	Top-3 TREC Runs Fused	0.380	0.572	0.459

Table 5.6: A recreation of Table 4.4 on page 111 to detail the effectiveness details of each of the challengers and champion system experiments explored in Chapter 4.

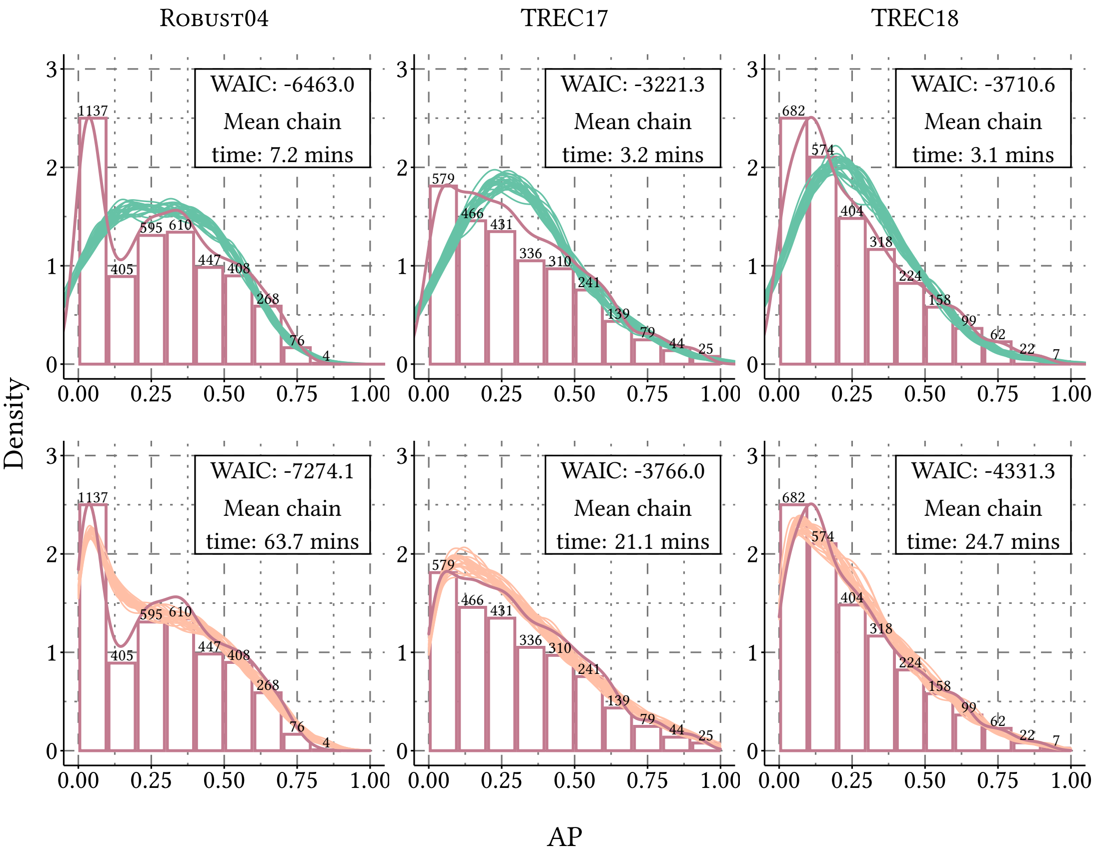


Figure 5.10: Draws from the posterior predictive distribution (thin lines) from the BHM-Gaussian statistical model (top row) against a BHM-ZOiB model (bottom row) for the AP metric on the three test collections studied in Chapter 4. The original score distribution for each test collection is represented with a pink density plot and accompanying histogram; identical for each vertical pairing. Where the PPD draws are closest to the original distribution, this indicates good model fit (supplemented with WAIC for a quantitative value). Note that the left-most column for Robust04 directly corresponds to the example presented in Figure 5.1 on page 128; the top-left graph presents 25 PPD draws instead of one. The mean wall time per chain for MCMC is also reported for each model.

In criticism of both models, neither captures the trough between the modes. However, the BHM-ZOiB model is consistently better than the BHM-Gaussian model on WAIC. Perfection is an unrealistic goal when attempting to model observed data with mathematical objects, and a key principle of Bayesian inference is the acceptance that more expressive models are likely become available in the future, particularly with advances in computing power. For the middle and right-most columns, TREC17 and TREC18 respectively, the BHM-Gaussian method fails to capture the bias towards low AP scores on both collections, whereas BHM-ZOiB captures the shape of the original score distributions well.

Although the BHM-ZOiB method appears to model IR scores with greater expressivity than the BHM-Gaussian approach, in an inferential context it is important to consider that both methods have reasonable agreement in determining system dominance on IR effectiveness scores without risk adjustment. The next subsection examines whether the risk transformation applied on PPDRisk$^-$ values computed using these different models affects the outcomes of significance, such that using more expensive models (BHM-ZOiB, for example) could be considered more appropriate if risk inference is an evaluation goal.

5.3.2 Results

How Risk Level Impacts PPDRisk$^-$ Power. Figure 5.9 on page 155 provided an optimistic overview of the potential for improvements in significance when larger r values are used in a multi-system setting ($r = 2$ vs. $r = 1$). Figure 5.11 on the next page combines visualizing the PPDRisk$^-$ distribution with revisiting the BRisk$^-$ experimental results presented in Figure 4.10 on page 120 over the Robust04, TREC17, and TREC18 collections.

In the previous experiments for $r = 1$, the BRisk$^-$ method was only able to separate the highly effective Challenger 4 system from the Champion system, as was indicated by the non-overlapping credibility intervals for the pairs of $\hat{\alpha}$ estimates. In Figure 5.11 the PPDRisk$^-$ outcome is the same as the above BRisk$^-$ Robust04 observation, interpreting significance as whether the predictive interval of the PPDRisk$^-$ estimate excludes zero. This pattern of only being able to distinguish the Challenger 4 system against the champion system continues in the TREC17 and TREC18 corpora for $r = 1$. When $r = 2$ and losses are given a two-fold weighting, both BRisk$^-$ and PPDRisk$^-$ for the Robust04 collection are unable (with 95% credibility) to determine whether any of the challengers are either more or less risky. Challenger 4 was able to retain significance when the score differences were doubled against the champion on TREC17 and TREC18.

The discrepancy in the characteristics of the PPDRisk$^-$ and BRisk$^-$ methodologies becomes clearer when $r = 5$ and $r = 10$. For the BRisk$^-$ method, no systems can be separated, and uncertainty appears to increase as r increases. However, for the PPDRisk$^-$ method, Challengers 1–3 are now considered significantly risky given a five-fold increase in loss weighting, and as r increases to $r = 10$, Challenger 4 also comes close to significant risk for Robust04. When $r = 10$ for TREC17 and TREC18, the four challengers are crossing over from indeterminate to significant, demonstrating how effective Challenger 4 is.

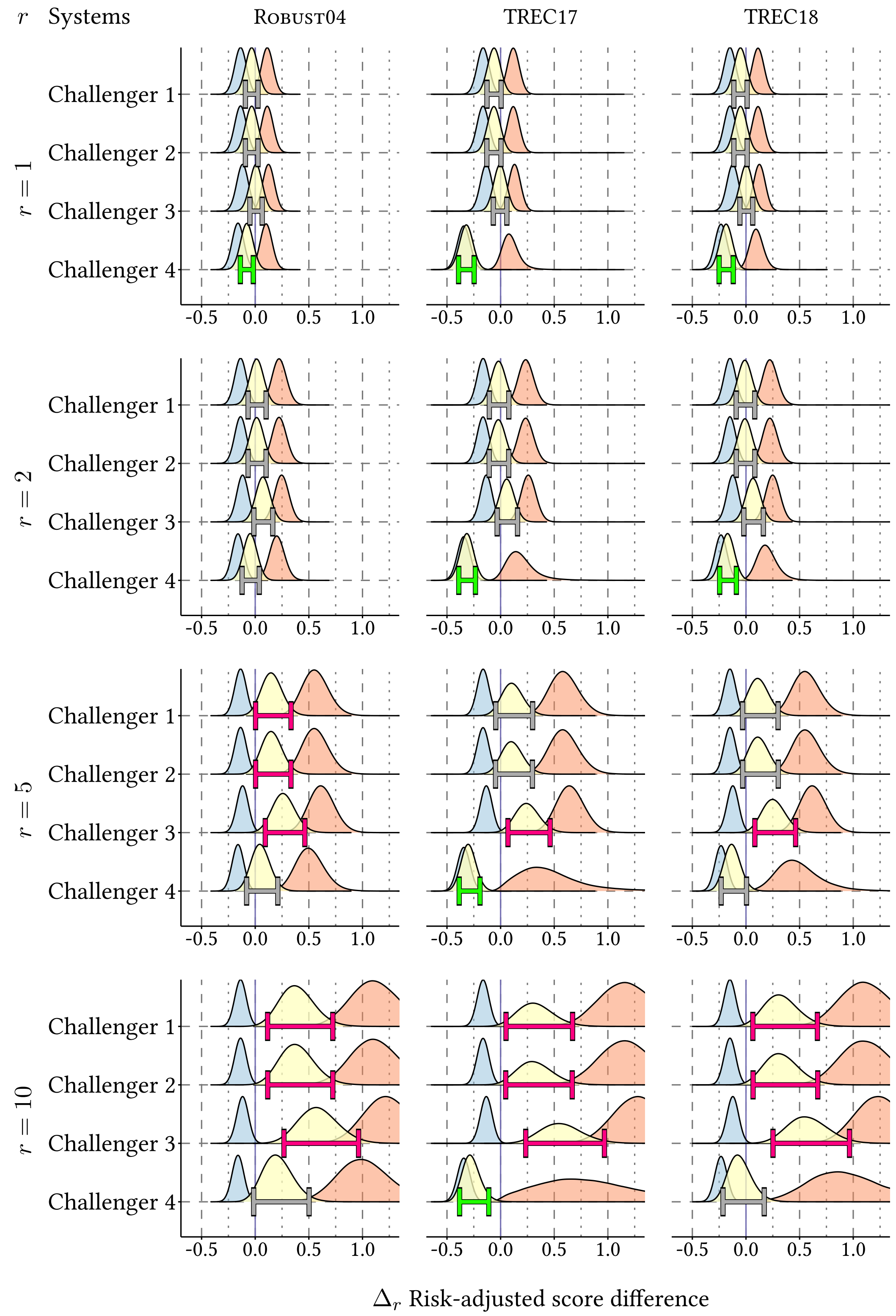


Figure 5.11: Champion system compared with risk-adjusted AP scores against four challengers using the PPDRisk$^-$ approach, for different risk levels on each of the collections. Here the deviation from normality in the risk score (shown to be problematic in previous chapters) is accounted for without explicitly adding further adjustments.

	Run		One vs. one				One vs. many (*a priori*)		One vs. many (*post hoc*)			
	System	AP	URisk$^-$	TRisk$^-$	BCa$^-$		BRisk$^-$		PPDRisk$^-$ (ZOiB)		PPDRisk$^-$ (Gaussian)	
Robust04	Champion	0.274	∅	∅	∅		−0.103	[−0.187, −0.021]	∅		∅	
	Challenger 1	0.323	−0.024	−1.408	−0.024	[−0.067, 0.020]	−0.122	[−0.206, −0.038]	0.152	[0.003, 0.334]	0.123	[−0.011, 0.286]
	Challenger 2	0.322	−0.026	−1.581	−0.026	[−0.066, 0.016]	−0.123	[−0.207, −0.039]	0.152	[0.004, 0.333]	0.124	[−0.010, 0.290]
	Challenger 3	0.264	0.105	2.976	0.105	[0.042, 0.236]	−0.024	[−0.107, 0.058]	0.262	[0.094, 0.463]	0.257	[0.096, 0.449]
	Challenger 4	0.380	−0.071	−2.345	−0.071	[−0.128, 0.031]	−0.156	[−0.241, −0.073]	0.047	[−0.080, 0.211]	0.015	[−0.095, 0.157]
TREC17	Champion	0.210	∅	∅	∅		0.017	[−0.065, 0.099]	∅		∅	
	Challenger 1	0.289	−0.052	−2.077	−0.052	[−0.100, 0.031]	−0.025	[−0.107, 0.057]	0.107	[−0.045, 0.298]	0.085	[−0.069, 0.277]
	Challenger 2	0.290	−0.053	−2.114	−0.053	[−0.100, 0.034]	−0.025	[−0.107, 0.057]	0.104	[−0.047, 0.297]	0.082	[−0.070, 0.276]
	Challenger 3	0.216	0.015	1.817	0.015	[−0.001, 0.043]	0.029	[−0.052, 0.110]	0.245	[0.069, 0.461]	0.269	[0.072, 0.503]
	Challenger 4	0.572	−0.352	−11.100	−0.352	[−0.419, −0.257]	−0.263	[−0.349, −0.179]	−0.301	[−0.386, −0.192]	−0.334	[−0.408, −0.247]
TREC18	Champion	0.236	∅	∅	∅		−0.116	[−0.207, −0.024]	∅		∅	
	Challenger 1	0.301	−0.040	−1.882	−0.040	[−0.091, 0.014]	−0.151	[−0.244, −0.060]	0.114	[−0.031, 0.299]	0.095	[−0.047, 0.274]
	Challenger 2	0.300	−0.042	−2.047	−0.042	[−0.092, 0.008]	−0.153	[−0.245, −0.061]	0.115	[−0.030, 0.301]	0.097	[−0.046, 0.275]
	Challenger 3	0.231	0.065	3.468	0.065	[0.027, 0.123]	−0.058	[−0.151, 0.033]	0.254	[0.083, 0.463]	0.270	[0.089, 0.487]
	Challenger 4	0.459	−0.165	−2.791	−0.165	[−0.266, 0.071]	−0.261	[−0.354, −0.169]	−0.131	[−0.232, 0.005]	−0.176	[−0.259, −0.071]

Table 5.7: A revisiting of the BRisk$^-$ experiment shown in Table 4.5, showing one-to-many Bayesian risk overlays. A fixed value of $r = 5$ is used; significant outcomes are shown in blue. Items in the BRisk$^-$ column are significant if the CI does not overlap with the Champion reference system.

PPDRisk$^-$ (BHM-Gaussian or BHM-ZOiB) Against BRisk$^-$. In exploring a more powerful approach than the BRisk$^-$ method proposed in Chapter 4, the results in this study include PPDRisk$^-$ as a new competing approach against other inferential risk methods in Table 5.7. ZRisk$^-$ and GeoRisk$^-$ are excluded as their inferential outcomes are not comparable. An advantage that the PPDRisk$^-$ method has over BRisk$^-$ is the ability to more easily update the generative statistical model for the effectiveness metric used. If Bayesian inferential modeling becomes more popular in IR and there are further efforts to fine tune the model employed for IR scores of a particular kind, it is more likely that improving how IR effectiveness scores are modeled will be the direct line of inquiry, rather than how to model risk-adjusted ones (if the PPDRisk$^-$ method proves to yield similar or better outcomes). For that reason, two PPDRisk$^-$ columns are provided in the table for comparing the relative outcomes of PPDRisk$^-$ using a BHM-Gaussian or BHM-ZOiB statistical model.

Since a champion-challenger pair is only significant for BRisk$^-$ if their credible intervals are non-overlapping, and PPDRisk$^-$ operates in the more traditional way where a predictive interval excluding zero is significant, the highlighted regions aid in identifying the salient details. When $r = 5$, PPDRisk$^-$ with either generative model dominates the BRisk$^-$ method in terms of statistical power. In the case of the only significant difference identified by BRisk$^-$ (Challenger 4 vs. Champion on TREC17), both PPDRisk$^-$ models agree with this assessment, as do all alternative frequentist approaches, including BCa$^-$ with Bonferroni correction applied. With that, in all cases where BCa$^-$ with Bonferroni correction has identified a significant difference, both PPDRisk$^-$ columns agree.

The disagreement in risk assessment when the evidence to reject the nulls are weak is interesting between the BHM-ZOiB and BHM-Gaussian generative models and indicates that the selected model plays a vital role in determining the outcome of the test. The BHM-ZOiB PPDRisk$^-$ combination has more findings of significance, but is biased towards finding more cases where a challenger is risky against the champion as evidenced by the first two rows. However, on the last row of the table, the BHM-Gaussian PPDRisk$^-$ combination finds that the Champion 4 system has no significant chance of harm against the champion for $r = 5$, where the BHM-ZOiB model is more reserved with its assessment.

5.3.3 Discussion

The above results indicate that PPDRisk$^-$ is at least as powerful as BRisk$^-$ for lower levels of r, and is better able to distinguish items for larger values of r (consistent with the expectations set from existing inferential risk metrics). Additionally, the BHM-ZOiB model works well with AP scores, and appears to follow the original data more closely than the BHM-Gaussian approach. It is conjectured that this more expressive model results in more careful decision making when used in combination with the PPDRisk$^-$ method, as it also happens to take the more conservative option of determining whether a system is risky or has no chance of harm for a given risk-level. This section therefore answers S-RQ5.3: *How does the sensitivity of posterior predictive risk compare with modeling risk-adjusted scores directly?*

5.4 Conclusion

It was previously unknown whether a powerful solution to performing risk inference in a Bayesian way existed using the sampling technologies available today. Although the BRisk^{-} method accounted for the desired properties of multiple comparison correction and acknowledges the asymmetrical shape of the risk-adjusted score distribution, its lack of power (especially for larger r values) is a key barrier to adoption. Because of this issue, this chapter explores whether there is an alternative way to compute risk inference using the PPD, while keeping the model of standard IR effectiveness values intact (RQ5). A logical consequence of that goal is to honor the distributional shape of IR effectiveness scores, and how risk inferences are impacted downstream when using the PPD on alternative Bayesian IR models.

This chapter started by improving the collective understanding of how more bespoke models might yield more statistical power when attempting to infer system differences. A Zero-One inflated Beta distribution (BHM-ZOiB) model was used as a more bespoke (yet expensive) contrast to the traditional BHM-Gaussian model used for IR scores. As Carterette [42] indicates, summing the document gain scores (whether it be RBP or otherwise) over each topic to obtain a single topic score results in losing information that could be exploited for improving inferential power. To that end, BHM-ZOiB-Rank is also compared in the system dominance exploration. The first round of the TREC COVID dataset was used as a recent modern ad-hoc search task, which features fewer topics and judgments than is typical of an offline evaluation campaign, making it a reasonable collection to use as a test-bed for exploring more expensive models as a base step before interpreting their feasibility on larger collections. The sub-research question S-RQ5.1 is answered in the affirmative, showing that more expensive models result in tighter credible intervals for the BHM-ZOiB model and high agreement with the BHM-Gaussian model regarding relative system dominance. The BHM-ZOiB-Rank method yielded tighter credible intervals than the above, but its relative ranking of systems also led to a degree of skepticism. Understanding why this was the case is noted as an area for future work in Section 6.2.1 on page 167, and modeling at the document-rank level rather than inferring over topic aggregates may be key to improving inferential power.

Section 5.2 detailed the novel PPDRisk^{-} method using the PPD to calculate risk using a worked example. The new PPDRisk^{-} method was empirically verified to provide practically useful inferences on the TREC COVID first round dataset, answering S-RQ5.2. Section 5.3 revisited the BRisk^{-} experiments from Chapter 4 to explore whether PPDRisk^{-} produces greater inferential power than BRisk^{-}. The BHM-ZOiB and BHM-Gaussian statistical models were both used, and both demonstrated that using the PPD for computing risk inference rather than attempting to model risk directly has more discriminative ability, answering S-RQ5.3. Inferential risk outcomes using PPDRisk^{-} did differ between BHM-ZOiB and BHM-Gaussian, particularly for larger values of r. More accurate models of the true score distribution will provide more precise risk inferences. However, in any situation requiring inference, oracular knowledge of whether an inference is valid does not exist.

6

Conclusion and Future Work

The interaction between IR evaluation and effectiveness research is symbiotic. Improving how systems are evaluated enables practitioners to have greater certainty that their retrieval models are performing well at the task of sifting through billions of documents in milliseconds, while improvements in system effectiveness help to calibrate evaluation goals, accuracy, and expectations. The well-established insight that topic performance varies substantially over systems plays an vital role in search effectiveness. That variance can potentially deceive practitioners into selecting systems that perform well on average, without considering the deleterious effect of a small but significant proportion of supported information needs. Even if the updated ranker is better on average, searchers may perceive it as less effective in general if topics they previously issued queries for are now ineffective.

Risk overlays have been proposed to address the problem above. However, many questions arise concerning inferentially evaluating risk comparisons between an existing system when there are many others to choose from, primarily when multiple comparisons should be corrected. How do the distributional properties of risk-adjusted scores impact the results of parametric statistical tests? How can risk-adjusted scores be modeled over multiple systems with multiple comparison correction? How can standard IR scores be modeled over multiple systems (with multiple comparison correction) to improve the sensitivity of multiple system risk inference? This thesis provides answers to these questions: the results of Chapter 3 found that nonparametric and parametric tests gave disparate outcomes on risk-adjusted scores, informing the inferential risk measure BRisk^{-} proposed in Chapter 4. The sensitivity of BRisk^{-} subsequently motivated exploring a more flexible approach in Chapter 5 named PPDRisk^{-}, yielding a more sensitive and practically useful inferential measure.

The above research outcomes described in this thesis unlock new opportunities for advancing the state-of-the-art in IR evaluation research, particularly in the Bayesian inference domain. Many practitioners are skeptical of the advantages of statistical inference in IR, where a common complaint is that a sample may not accurately represent the population of topics. Bayesian modeling enables practitioners to not only make inferences about what was observed, it allows for adjusting modeling beliefs over time with clean solutions to the traditional problems: score volatility, effect sizes, multiple comparison correction, and deviations from normality.

6.1 Thesis Outcomes

Firstly, Chapter 3 investigated possible improvements to risk inference in the foundational paired system setting. The key motivation behind this exploration was to validate the modeling assumptions in the paired scenario, and evaluate any possible impacts they may have when proposing an extension to multiple system risk testing with correction. The merit of a smooth-value risk function was also explored in contrast to the outcomes of the canonical piece-wise approach. Pioneering risk inference techniques in IR assume scores can be modeled using a t-distribution in the case of TRisk$^-$, and multiple system inference is conducted against a synthetic run built by aggregating the scores of a collection of runs (ZRisk$^-$ and GeoRisk$^-$). Since the loss-scaling part of the score differences distribution is one-sided and null hypothesis statistical tests like the Student t-test assume symmetry, it is explored whether deviations from normality affect the outcomes of t-distributed risk inferences. An empirical analysis comparing the confidence intervals produced using parametric and non-parametric testing approaches revealed that larger values of r tend to result in over-confident CIs in the TRisk$^-$ case. The need to correct the confidence intervals for the resulting bias in the risk-adjusted distribution is the first contribution of this thesis, with BCa$^-$ as a proposed solution to testing in the paired case.

Using the insight that the skewness of the risk-adjusted score difference distributions are a factor in the interpretation of inferences, Chapter 4 introduces a method whereby Bayesian modeling can be employed to test risk scores. Pairs of systems at a time are modeled parametrically using a three-parameter skew-normal distribution, which is extended to a multiple system testing methodology. An innovative inferential evaluation scenario is presented with the goal of involving a champion system, multiple challengers, and the entire set of known submitted runs to an offline evaluation track in the testing outcomes. The power of the proposed multiple system BRisk$^-$ approach is compared in context with frequentist risk tests (including BCa$^-$), with and without correcting the false discovery rate. The operating characteristics of BRisk$^-$ were consistent with other frequentist tests, yielding a conservative testing mechanism (albeit too conservative for practical use). As larger values of r are used with BRisk$^-$, it became more difficult to achieve statistical significance, forming the basis for re-exploring the problem differently in the subsequent chapter. This chapter contributed the first application of Bayesian hierarchical modeling for multiple system inference with correction, and laid the foundations motivating the potential for Bayesian modeling on risk scores.

Finally, Chapter 5 demonstrates that flipping the modeling problem to Bayesian modeling of traditional IR scores reveals a more powerful approach to performing risk inference, using the posterior predictive distribution. Analysis of the TREC COVID track results is used to justify that including score uncertainty when interpreting traditional IR scores is as important as it ever was, especially in the context of collections with shallow judgments. The insight that the posterior predictive distribution can be used to form images of the original data within statistical expectations formed the basis for a new multiple system risk inference methodology PPDRisk^{-}, breaking from the tradition of modeling risk-adjusted scores directly, and instead treating it as a summary statistic. The experiments in Chapter 4 are re-evaluated with PPDRisk^{-} with the results indicating that independent of the set of priors used to establish the posterior of the typical IR scores, using PPDRisk^{-} yields a demonstrably more powerful testing approach than BRisk^{-} and gives consistent results with other testing mechanisms. Further, since typical IR scores have already undergone multiple comparison correction, PPDRisk^{-} scores are also transformed, and the skewness discussed in Chapter 3 is already accounted for in the generated credible intervals. This chapter contributes the first practical multiple system risk testing tool PPDRisk^{-}; achieving the overall goal of this thesis.

These three contributions represent an important step forward for conducting risk inference analyses of IR systems, in pairs, and now over multiple systems with correction. The research conducted in this thesis involving Bayesian modeling on IR effectiveness scores has brought to light many opportunities for further research. It is hoped that the dissemination of this thesis will encourage research in this area, and the tools and techniques presented here will reduce its barriers to entry. However, Bayesian inferences require more investment to understand than traditional statistical tests, and are far more computationally expensive than their frequentist counterparts. The following section discusses ways in which these ideas could be further extended.

6.2 Future Work

6.2.1 Modeling Missing Information

Carterette [42] first introduced the idea of inferring system dominance via document relevance in a system-topic ranking, where Chapter 5 explored modeling RBP gain scores through the novel ZOiB-Rank approach. Inference, in this way, can provide more sophisticated credibility intervals between systems with unjudged documents and fully pooled ones. For example, an RBP score is a vector quantity composed of the minimum RBP score and the residual error, where the residual error informs the practitioner what the RBP score could have been if the unjudged documents in the ranking were relevant. One of the key benefits of using Bayesian inference is the ability to handle unbalanced study designs [141], meaning that documents without judgments across different topics can be treated as missing values, rather than assuming irrelevance and assigning a document gain of zero. That is, the uncertainty in the score information (RBP residual) can be integrated into the credible intervals for system inference in combination with topic effects to capture a more comprehensive inference than is currently possible. How extreme uncertainty affects the performance of MCMC sampling and to what extent inferences can be made with incomplete information deserves a more thor-

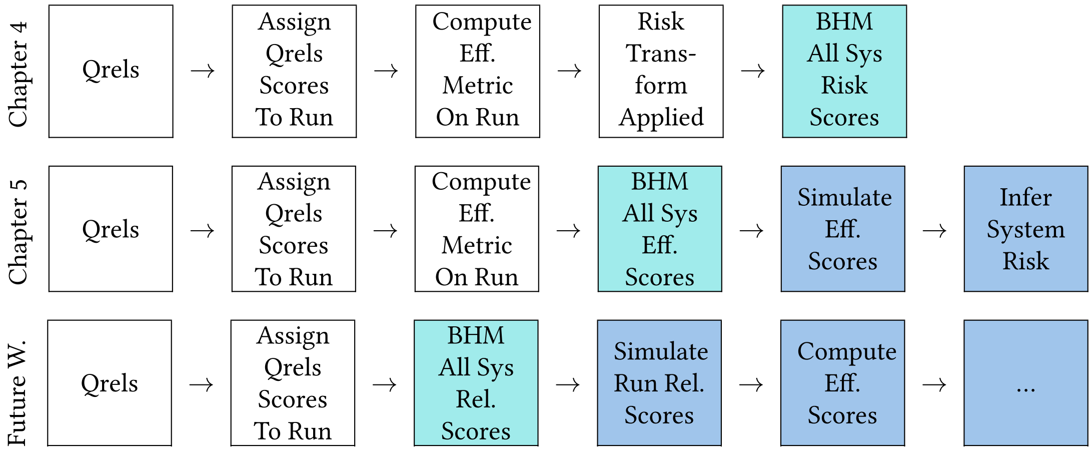


Figure 6.1: An overview of the Bayesian hierarchical modeling (BHM) experiments explored in Chapter 4 and Chapter 5 in terms of performing risk inference on many systems for a given metric. An exciting avenue of future work would be to apply the Carterette [42] idea of modeling relevance directly, to avoid the need for separate models for each metric, while applying the family-wise error correction approach of BHM to be able to infer over every system of interest. Boxes marked in cyan indicate where the Bayesian model is simulated, and blue boxes refer to where simulations from the Bayesian model are used to answer different questions about the data.

ough investigation, given the number of different MCMC sampling algorithms that have been proposed and their pros and cons concerning hierarchically modeled IR scores. Observing the RBP residual remains an important diagnostic in the trustworthiness of an effectiveness score.

6.2.2 Modeling Relevance Directly Over Many Systems

Section 6.2.1 discussed future work exploring the potential of modeling the per-document RBP gain score contributions while labeling documents with unknown relevance values as unknown, rather than zero, to improve the quality of the model induced. Carterette [42] had previously advocated modeling relevance scores directly rather than scores tied to a particular metric. However, not much work had been done around the topic, and it was unclear what the benefits might be for statistical power. The ZOiB-Rank experiments in Chapter 5 demonstrate promise in achieving tighter confidence intervals for the system effects by retaining relevance details over topics instead of aggregating them into one score.

Figure 6.1 recaps the modeling scenarios encountered in this thesis with the future work idea of further exploring modeling relevance directly. Chapter 4 explored the idea of using Bayesian hierarchical modeling over many systems at once, and the results indicate that this approach was possible, but modeling risk itself was not particularly useful as it encapsulated too much noise. Chapter 5 modeled IR effectiveness scores directly and used the PPD to construct predictive intervals for the risk values, resulting in more statistical power. However, many systems could not be statistically separated from the best or worst in the set. The bottom row in Figure 6.1 describes a possible future scenario where the relevance values in a run are modeled and simulated many times to explore the metrics' volatility without

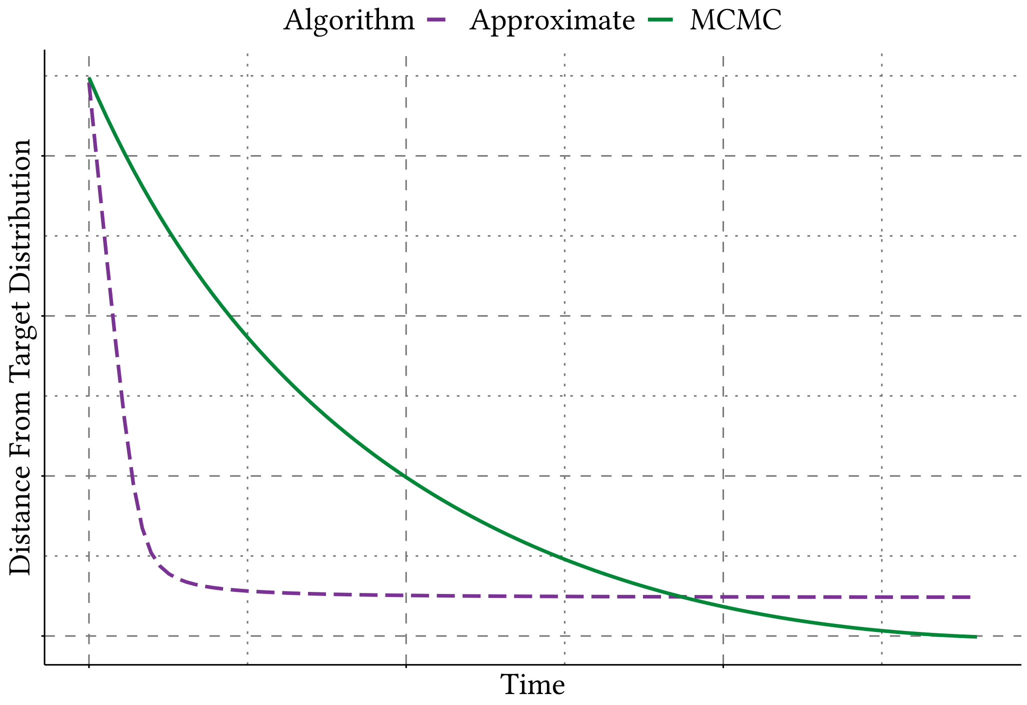


Figure 6.2: An adaptation of the sketch by Gelman et al. [90, Figure 5], showing the potential efficiency improvements of using approximate posterior samplers instead of MCMC.

needing to model each metric explicitly. Recall that the skewness was implicitly accounted for in the risk scores when calculated via the PPD in Chapter 5. Although IR practitioners know that multiple testing across systems without correction is bad practice, using multiple metrics until a meaningful outcome is achieved is not methodologically sound either [212]. (The experiments in this thesis were careful to use a small set of metrics to model system effectiveness, as each new model induced by a metric increases the likelihood of detecting a significant difference erroneously.) By modeling relevance directly, the same simulated set of relevance values for a system-topic combination can be issued to many rankers, mitigating against multiple comparison issues and reducing the impact of the look-elsewhere effect.

6.2.3 More Efficient Bayesian Inference

An important point of differentiation between classical and Bayesian statistics is the amount of computation required to simulate the posterior. Many MCMC algorithms exist, and like most algorithmic problems, there are trade-offs between efficiency and accuracy in their implementations. For example, when the volume preservation property of the sampler is relaxed, larger efficiency gains are possible:

> "While all other samplers return 5,000 samples in less than 20 minutes, the baseline NUTS could not draw more than 2,000 samples after 24 hours."
>
> — Afshar et al. [6, p. 8]

While MCMC methods are guaranteed to produce asymptotically exact samples from the target density [29], *variational inference* (VI) is another class of probabilistic sampler with relaxed properties to improve efficiency. Figure 6.2 on the previous page provides a visual comparison of MCMC and VI, adapted from Gelman et al. [90, Figure 5]. How amenable this algorithm class is to hierarchically modeled IR data and the potential for sampling error is worthy of investigation to improve inference speeds. If the assumptions of the NUTS MCMC sampler cannot be relaxed, Tran et al. [188] show that it is possible to achieve a $100\times$ speed-up against Stan's NUTS sampler with a multi-GPU architecture. Improving the efficiency of Bayesian sampling is important for situations where the model needs to be re-simulated frequently on new data, or where a complicated model is used on a large dataset.

Bibliography

[1] The ClueWeb09 dataset: Dataset details. http://lemurproject.org/clueweb09/index.php#Specs. (Accessed on 19th May 2022).

[2] Bayesian predictions. https://www.stata.com/features/overview/bayesian-predictions/#ppvalues. (Accessed on 17th February 2021). A tutorial on how posterior predictive values can be used in the Stata statistical tool.

[3] TREC Washington Post corpus. https://trec.nist.gov/data/wapost/. (Accessed on 19th May 2022).

[4] Library of congress classification. https://www.loc.gov/catdir/cpso/lcc.html, January 2014. (Accessed on 24th November 2022).

[5] M. Abdellaoui and E. Kemel. Eliciting prospect theory when consequences are measured in time units: "Time is not money". *Manag. Sci.*, 60(7):1844–1859, 2013.

[6] H. M. Afshar, R. Oliveira, and S. Cripps. Non-volume preserving Hamiltonian Monte Carlo and No-U-Turn Samplers. In *Proc. Int. Conf. on Artif. Intell. Stat. (AISTATS)*, pages 1675–1683, 2021.

[7] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *J. of the American Society for Information Science and Technology*, 61(5):859–868, 2010.

[8] A. Al-Maskari and M. Sanderson. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management*, 47(5):719–729, 2011.

[9] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 773–774, 2007.

[10] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: Does the test collection predict users' effectiveness? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 59–66, 2008.

[11] A. O. Alanazi, M. Sanderson, Z. Bao, and J. Kim. The impact of ad quality and position on mobile SERPs. In *Proc. ACM Conf. on Human Information Interaction and Retrieval (CHIIR)*, pages 318–322, 2020.

[12] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 433–440, 2005.

[13] J. Allan, D. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees. TREC 2017 common core track overview. In *Proc. Text Retrieval Conf. (TREC)*, pages 1–14, 2017.

[14] J. Arguello, F. Diaz, J. Lin, and A. Trotman. SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1147–1148, 2015.

[15] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 601–610, 2009.

[16] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian J. of Statistics*, 12(2):171–178, 1985.

[17] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 625–634, 2015.

[18] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 395–404, 2017.

[19] D. Banks, P. Over, and N.-F. Zhang. Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, 1(1):7–34, 1999.

[20] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect of multiple query variations on information retrieval system performance. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 339–346, 1993.

[21] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448, 1995.

[22] R. Benham and J. S. Culpepper. Risk-reward trade-offs in rank fusion. In *Proc. Australasian Document Computing Symp. (ADCS)*, pages 1:1–1:8, 2017.

[23] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. Mackenzie. Towards efficient and effective query variant generation. In *Proc. Conf. on Design of Experimental Search & Information Retrieval Systems (DESIRES)*, 2018.

[24] R. Benham, J. Mackenzie, A. Moffat, and J. S. Culpepper. Boosting search performance using query variations. *ACM Trans. on Information Systems*, 37(4):41, 2019.

[25] R. Benham, B. Carterette, J. S. Culpepper, and A. Moffat. Bayesian inferential risk evaluation on multiple IR systems. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 339–348, 2020.

[26] R. Benham, A. Moffat, and J. S. Culpepper. Bayesian system inference on shallow pools. In *Proc. European Conf. on Information Retrieval (ECIR)*, pages 209–215, 2021.

[27] J. Berkhof, I. Van Mechelen, and H. Hoijtink. Posterior predictive checks: Principles and discussion. *Computational Statistics*, 15(3):337–354, 2000.

[28] L. Berul. Information storage and retrieval, a state-of-the-art report. `https://apps.dtic.mil/sti/citations/AD0630089`, 1964. (Accessed on 25th November 2022).

[29] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *J. of the American Statistical Association*, 112(518):859–877, 2017.

[30] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936. In Italian, cited via Weisstein [214].

[31] J. Boyan, D. Freitag, and T. Joachims. A machine learning architecture for optimizing web search engines. In *AAAI Workshop Internet Based Inf. Syst.*, pages 1–8, 1996.

[32] L. Boytsov, A. Belova, and P. Westfall. Deciding on an adjustment for multiplicity in IR experiments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 403–412, 2013.

[33] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[34] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 33–40, 2000.

[35] C. Buckley and E. M. Voorhees. Retrieval system evaluation. *TREC: Experiment and Evaluation in Information Retrieval*, pages 53–75, 2005. Chapter 3.

[36] P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *J. of Statistical Software*, 80:1–28, 2017.

[37] P.-C. Bürkner. Parameterization of response distributions in brms, 2022. URL `https://cran.r-project.org/web/packages/brms/vignettes/brms_families.html#location-shift-models`. (Accessed on 27th October 2022).

[38] M. K. Cain, Z. Zhang, and K-.H. Yuan. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, 49(5):1716–1735, 2017.

[39] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *J. of Statistical Software*, 76(1), 2017.

[40] B. Carterette. Model-based inference about IR systems. In *Proc. Int. Conf. on Theory of Information Retrieval (ICTIR)*, pages 101–112, 2011.

[41] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. on Information Systems*, 30(1):4, 2012.

[42] B. Carterette. Bayesian inference for information retrieval evaluation. In *Proc. Int. Conf. on Theory of Information Retrieval (ICTIR)*, pages 31–40, 2015.

[43] B. Carterette and M. D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 643–652, 2007.

[44] C. Castillo. Fairness and transparency in ranking. *SIGIR Forum*, 52(2):64–71, 2019.

[45] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 621–630, 2009.

[46] A. Checco, K. Roitero, E. Maddalena, S. Mizzaro, and G. Demartini. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proc. Conf. Human Computation and Crowdsourcing (HCOMP)*, pages 11–20, 2017.

[47] H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 429–436, 2006.

[48] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 web track. In *Proc. Text Retrieval Conf. (TREC)*, pages 1–8, 2012.

[49] C. W. Cleverdon. The evaluation of systems used in information retrieval. In *Proceedings of the international conference on scientific information*, volume 1, pages 687–698. National Academy of Sciences Washington, DC„ 1959.

[50] C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib proceedings*, pages 173–192, 1967.

[51] C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 3–12, 1991.

[52] J. Cohen. *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates, New York, NY, USA, second edition, 1988.

[53] K. Collins-Thompson. Accounting for stability of retrieval algorithms using risk-reward curves. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 27–28, 2009.

[54] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 837–846, 2009.

[55] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2013 web track overview. In *Proc. Text Retrieval Conf. (TREC)*, 2014.

[56] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2014 web track overview. In *Proc. Text Retrieval Conf. (TREC)*, 2015.

[57] D. Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3):140216, 2014.

[58] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 533–540, 2006.

[59] G. V. Cormack, C. L. A. Clarke, and S. Büttcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 758–759, 2009.

[60] C. A. Cuadra and R. V. Katter. Opening the black box of 'relevance'. *J. of Documentation*, 1967.

[61] J. S. Culpepper, G. Faggioli, N. Ferro, and O. Kurland. Topic difficulty: Collection and query formulation effects. *ACM Trans. on Information Systems*, 40(1):19, 2021.

[62] A. C. Davison and D. B. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge University Press, Cambridge, UK, 1997.

[63] M. Dewey. *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library.* Amherst, Mass., 1876.

[64] M. Dietze. Bayesian hierarchical models, 2020. URL `https://www.youtube.com/watch?v=SMWleVKO9ZM`. (Watched on 17th May 2021).

[65] B. T. Dinçer, C. Macdonald, and I. Ounis. Risk-sensitive evaluation and learning to rank using multiple baselines. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 483–492, 2016.

[66] B. Dinçer, I. Ounis, and C. Macdonald. Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In *Proc. European Conf. on Information Retrieval (ECIR)*, pages 26–38, 2014.

[67] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 23–32, 2014.

[68] G. Dupret and B. Piwowarski. A user behavior model for average precision and its generalization to graded judgments. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pages 531–538, 2010.

[69] B. Efron. Better Bootstrap confidence intervals. *J. of the American Statistical Association*, 82(397):171–185, 1987.

[70] E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *J. of Financ. Econ.*, 33(1):3–56, 1993.

[71] E. F. Fama and K. R. French. A five-factor asset pricing model. *J. of Financ. Econ.*, 116 (1):1–22, 2015.

[72] N. Ferro and M. Sanderson. Sub-corpora impact on system effectiveness. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 901–904, 2017.

[73] N. Ferro and M. Sanderson. How do you test a test? A multifaceted examination of significance tests. In *Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pages 280–288, 2022.

[74] N. Ferro and G. Silvello. A general linear mixed models approach to study system component effects. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 25–34, 2016.

[75] N. Ferro, Y. Kim, and M. Sanderson. Using collection shards to study retrieval performance effect sizes. *ACM Trans. on Information Systems*, 37(3):30, 2019.

[76] C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical distributions.* John Wiley & Sons, Hoboken, NJ, USA, fourth edition, 2011.

[77] E. A. Fox and J. A. Shaw. Combination of multiple searches. *Proc. Text Retrieval Conf. (TREC)*, pages 243–252, 1994.

[78] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. on Information Systems*, 23(2):147–168, 2005.

[79] M. Franke and T. Roettger. Bayesian regression modeling (for factorial designs): A tutorial, July 2019. (Accessed on 20th May 2022).

[80] E. Frøkjær, M. Hertzum, and K. Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)*, pages 345–352, 2000.

[81] L. Gallagher, J. Mackenzie, and J. S. Culpepper. Revisiting spam filtering in web search. In *Proc. Australasian Document Computing Symp. (ADCS)*, pages 1–4, 2018.

[82] A. Gelman. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electron. J. Statist.*, 7:2595–2602, 2013.

[83] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, 2013. URL `http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf`. (Accessed on 19th August 2022).

[84] A. Gelman and I. Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, 2006.

[85] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Stat. Sci*, 7(4):457–472, 1992.

[86] A. Gelman and C. R. Shalizi. Philosophy and the practice of Bayesian statistics. *British J. of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

[87] A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760, 1996.

[88] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *J. Res. Int. Educ.*, 5(2):189–211, 2012.

[89] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis.* Chapman and Hall/CRC, Boca Raton, Florida, USA, 2013.

[90] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint*, abs/2011.01808, 2020.

[91] D. George and P. Mallery. *SPSS for Windows step by step: A simple study guide and reference, 17.0 update, 10/e.* Pearson Education, Bengaluru, India, 2011.

[92] G. V. Glass, P. D. Peckham, and J. R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Res. Educ.*, 42(3):237–288, 1972.

[93] G. Hamra, R. MacLehose, and D. Richardson. Markov Chain Monte Carlo: An introduction for epidemiologists. *Int. J. of Epidemiol.*, 42(2):627–634, 2013.

[94] D. Harman. Overview of the first TREC conference. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 36–47, 1993.

[95] S. H. Hashemi and J. Kamps. University of Amsterdam at TREC 2014: Contextual suggestion and web tracks. In *Proc. Text Retrieval Conf. (TREC)*, 2014.

[96] D. Hawking and N. Craswell. Overview of the TREC-2001 web track. In *Proc. Text Retrieval Conf. (TREC)*, pages 61–67, 2002.

[97] D. Hawking and P. Thistlewaite. Overview of TREC-6 very large collection track. In *Proc. Text Retrieval Conf. (TREC)*, pages 93–106, 1998.

[98] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3):1–15, 2015.

[99] W. Hersh, A. Turpin, S. Price, B. Chanjamin, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 17–24, 2000.

[100] T. Hesterberg, S. Monaghan, D. S. Moore, A. Clipson, and R. Epstein. *Companion chapter 18: Bootstrap methods and permutation tests for the practice of business statistics.* WH Freeman & Co., New York, NY, USA, 2003.

[101] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res*, 15(1):1593–1623, 2014.

[102] K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval. *Foundations & Trends in Information Retrieval*, 10(1):1–117, 2016.

[103] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, pages 65–70, 1979.

[104] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 329–338, 1993.

[105] J. Jiang, D. He, D. Kelly, and J. Allan. Understanding ephemeral state of relevance. In *Proc. ACM Conf. on Human Information Interaction and Retrieval (CHIIR)*, pages 137–146, 2017.

[106] D. N. Joanes and C. A. Gill. Comparing measures of sample skewness and kurtosis. *J. of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189, 1998.

[107] K. S. Jones. Index term weighting. *Information Storage and Retrieval*, 9(11):619–633, 1973.

[108] K. S. Jones. *Information retrieval experiment.* Butterworth-Heinemann, Oxford, UK, 1981.

[109] K. S. Jones and C. Van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. *British Library Research and Development Report 5553*, 1975. Cited via Voorhees and Harman [204]. (Accessed on 19th August 2022).

[110] K. S. Jones and P. Willett. *Readings in information retrieval.* Morgan Kaufmann, San Francisco, CA, USA, 1997.

[111] H.-Y. Kim. Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, 38(1):52, 2013.

[112] J. Kirakowski and M. Corbett. SUMI: The software usability measurement inventory. *British J. of Educational Technology*, 24(3):210–2, 1993.

[113] C. M. R. Kitchen. Nonparametric vs parametric tests of location in biomedical research. *American J. of Ophthalmology*, 147(4):571–572, 2009.

[114] A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933. In Italian, cited via Razali and Wah [150].

[115] J. K. Kruschke. Bayesian estimation supersedes the t-test. *J. of Experimental Psychology: General*, 142(2):573, 2013.

[116] J. K. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press, London, UK, 2014.

[117] J. K. Kruschke and T. M. Liddell. The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.*, 25(1):178–206, 2018.

[118] B. Lambert. *A Student's guide to Bayesian statistics.* Sage, Newbury Park, CA, USA, 2018.

[119] C. Liu, X. Yan, and J. Han. Mining control flow abnormality for logic error isolation. In *Proc. SIAM Int. Conf. on Data Mining (SDM)*, pages 106–117, 2006.

[120] T.-Y. Liu. Learning to rank for information retrieval. *Foundations & Trends in Information Retrieval*, 3(3):225–331, 2009.

[121] P. Loewerre and J. Dominiquini. Overcoming the barriers to effective innovation. *Strategy & Leadership*, 34(1):24–31, 2006.

[122] X. Lu. *Efficient and effective retrieval using higher-order proximity models*. PhD thesis, RMIT University, 2017. URL `https://researchrepository.rmit.edu.au/esploro/outputs/doctoral/Efficient-and-effective-retrieval-using-Higher-Order-proximity-models/9921863900501341`. (Accessed on 29th September 2022).

[123] X. Lu, A. Moffat, and J. S. Culpepper. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval*, 19(4):416–445, 2016.

[124] M. Maiti. A critical review on evolution of risk factors and factor models. *J. of Econ. Surveys*, 34(1):175–184, 2020.

[125] L. Maligranda. The AM-GM inequality is equivalent to the Bernoulli inequality. *Math. Intell.*, 34(1):1–2, 2012.

[126] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 267–275, 2001.

[127] J. Manotumruksa, C. Macdonald, and I. Ounis. A contextual recurrent collaborative filtering framework for modelling sequences of venue checkins. *Information Processing & Management*, 57(6):102092, 2020.

[128] E. Markowitz. Portfolio selection. *J. of Financ.*, 7(1):77–91, 1952.

[129] M. Matsumoto and T. Nishimura. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. on Model. Comput. Simul..*, 8(1):3–30, 1998.

[130] R. McCreadie, R. Deveaud, M. Albakour, S. Mackie, C. Macdonald, I. Ounis, T. Thonet, and B. T. Dinçer. University of Glasgow at TREC 2014: Experiments with Terrier in contextual suggestion, temporal summarisation and web tracks. In *Proc. Text Retrieval Conf. (TREC)*, 2014.

[131] J. H. McDonald. *Handbook of Biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009. URL `http://www.biostathandbook.com/wilcoxonsignedrank.html`.

[132] X.-L. Meng. Posterior predictive p-values. *Annals of Statistics*, 22(3):1142–1160, 1994.

[133] S. Mizzaro and S. Robertson. HITS hits TREC: Exploring IR evaluation results with network analysis. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 479–486, 2007.

[134] A. Moffat and A. F. Wicaksono. Users, adaptivity, and bad abandonment. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 897–900, 2018.

[135] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. on Information Systems*, 27(1):2, 2008.

[136] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 659–668, 2013.

[137] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled evaluation over query variations: Users are as diverse as systems. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1759–1762, 2015.

[138] C. N. Mooers. Mooers' law: Or, why some retrieval systems are used and others are not. *Technical Bull. 136*, 1959. Cited via Berul [28]. (Source nonrecoverable.).

[139] C. Muth, Z. Oravecz, and J. Gabry. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quant. Methods Psychol.*, 14(2):99–119, 2018.

[140] Y. Nagata. How to design the sample size. *Asakura Shoten*, 2003. In Japanese, cited via Sakai [160].

[141] L. Nalborczyk, C. Batailler, H. Lœvenbruck, A. Vilain, and P.-C. Bürkner. An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *J. of Speech, Language, and Hearing Research*, 62(5): 1225–1242, 2019.

[142] B. K. Nayak and A. Hazra. How to choose the right statistical test? *Indian J. of Ophthalmology*, 59(2):85, 2011.

[143] A. O'Hagan and T. Leonard. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1):201–203, 1976.

[144] R. Ospina and S. L. P. Ferrari. Inflated beta distributions. *Statistical Papers*, 51(1):111–126, 2010.

[145] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, and A. Barreiro. Using score distributions to compare statistical significance tests for information retrieval evaluation. *J. of the American Society for Information Science and Technology*, 71(1):98–113, 2019.

[146] M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proc. Int. Workshop on Dist. Stat. Comput. (DSC)*, 2003. ISSN 1609-395X.

[147] M. Plummer, A. Stukalov, and M. Denwood. rjags: Bayesian graphical models using MCMC, 2022. URL `https://cran.r-project.org/web/packages/rjags/rjags.pdf`. (Accessed on 11th November 2022).

[148] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 275–281, 1998.

[149] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[150] N. M. Razali and Y. B. Wah. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[151] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. Text Retrieval Conf. (TREC)*, 1994.

[152] J. J. Rocchio. Relevance feedback in information retrieval. *Information Storage and Retrieval Scientific Report No. ISR-9.*, 1965. (Accessed on 19th August 2022).

[153] P. H. S. Rodrigues, D. X. de Sousa, T. C. Rosa, and M. A. Gonçalves. Risk-sensitive deep neural learning to rank. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 803–813, 2022.

[154] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A review of modern computational algorithms for Bayesian optimal design. *Int. Stat. Rev.*, 84(1):128–154, 2016.

[155] F. Saitoaki, Y. Shoji, and Y. Yamamoto. Highlighting weasel sentences for promoting critical information seeking on the web. In *Int. Conf. on Web Information Systems Engineering (WISE)*, pages 424–440, 2020.

[156] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 525–532, 2006.

[157] T. Sakai. Designing test collections for comparing many systems. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 61–70, 2014.

[158] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

[159] T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 5–14, 2016.

[160] T. Sakai. Topic set size design. *Information Retrieval*, 19(3):256–283, 2016.

[161] T. Sakai. The probability that your hypothesis is correct, credible intervals, and effect sizes for IR evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 25–34, 2017.

[162] T. Sakai. *Laboratory experiments in information retrieval.* Springer, Singapore, 2018.

[163] T. Sakai and Z. Zeng. Which diversity evaluation measures are "good"? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 595–604, 2019.

[164] T. Sakai and Z. Zeng. Good evaluation measures based on document preferences. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 359–368, 2020.

[165] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *J. of the American Society for Information Science and Technology*, 41(4):288–297, 1990.

[166] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[167] G. Salton, E. A. Fox, and H. Wu. Extended Boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

[168] M. Sanderson and W. B. Croft. The history of information retrieval research. *Proc. IEEE*, 100 (Special Centennial Issue):1444–1451, 2012.

[169] M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 162–169, 2005.

[170] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 555–562, 2010.

[171] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. Differences in effectiveness across sub-collections. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1965–1969, 2012.

[172] M. P. Satija and J. Singh. Colon classification (cc). *Knowledge Organization*, 44(4):291–307, 2017.

[173] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965.

[174] W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *J. of Financ.*, 19(3):425–442, 1964.

[175] W. Sievers. Card catalog at Baillieu library, University of Melbourne, Victoria, 1961. URL `https://trove.nla.gov.au/work/26984502`. Photograph. (Accessed on 19th August 2022).

[176] C. L. Smith and P. B. Kantor. User adaptation: Good results from poor systems. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 147–154, 2008.

[177] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 95–104, 2012.

[178] M. D. Smucker and C. P. Jethani. Human performance and retrieval precision revisited. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 595–602, 2010.

[179] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 623–632, 2007.

[180] M. D. Smucker, J. Allan, and B. Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 630–631, 2009.

[181] Y. Song, X. Shi, R. White, and A. H. Awadallah. Context-aware web search abandonment prediction. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 93–102, 2014.

[182] D. X. D. Sousa, S. D. Canuto, T. C. Rosa, W. S. Martins, and M. A. Gonçalves. Incorporating risk-sensitiveness into feature selection for learning to rank. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 257–266, 2016.

[183] J. D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035, 2003.

[184] M. Taube, C. D. Gull, and I. S. Wachtel. Unit terms in coordinate indexing. *J. of the American Society for Information Science*, 3(4):213, 1952.

[185] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 94–101, 2006.

[186] P. Thomas, A. Moffat, P. Bailey, F. Scholer, and N. Craswell. Better effectiveness metrics for SERPs, cards, and rankings. In *Proc. Australasian Document Computing Symp. (ADCS)*, pages 1–8, 2018.

[187] H. Toyoda. Fundamentals of Bayesian statistics: Practical getting started by Hamiltonian Monte Carlo method. *Asakura Shoten*, 2015. In Japanese, cited via Sakai [161].

[188] D. Tran, M. W. Hoffman, D. Moore, C. Suter, S. Vasudevan, and A. Radul. Simple, distributed, and accelerated probabilistic programming. In *Proc. Conf. on Neural Information Processing Systems (NIPS)*, pages 7609–7620, 2018.

[189] J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2): 99–114, 1949.

[190] J. W. Tukey. *Exploratory data analysis*, volume 2. Pearson, New York, NY, USA, 1977.

[191] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 225–231, 2001.

[192] A. Turpin and W. Hersh. User interface effects in past batch versus user experiments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 431–432, 2002.

[193] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 11–18, 2006.

[194] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *J. of Risk and Uncertainty*, 5(4):297–323, 1992.

[195] J. Urbano and T. Nagler. Stochastic simulation of test collections: Evaluation scores. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 695–704, 2018.

[196] J. Urbano, M. Marrero, and D. Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 925–928, 2013.

[197] J. Urbano, H. Lima, and A. Hanjalic. Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 505–514, 2019.

[198] J. Urbano, M. Corsi, and A. Hanjalic. How do metric score distributions affect the type I error rate of statistical significance tests in information retrieval? In *Proc. Int. Conf. on Theory of Information Retrieval (ICTIR)*, pages 245–250, 2021.

[199] E. M. Voorhees. Overview of TREC 2004 robust retrieval track. In *Proc. Text Retrieval Conf. (TREC)*, pages 69–77, 2004.

[200] E. M. Voorhees. The TREC robust retrieval track. *SIGIR Forum*, 39(1):11–20, 2005.

[201] E. M. Voorhees. Effect on system rankings of further extending pools for TREC-COVID round 1 submissions, 2020. URL `https://ir.nist.gov/covidSubmit/papers/rnd1runs_j0.5-2.0.pdf`. Unpublished TREC Report. (Accessed on 19th August 2022).

[202] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 316–323, 2002.

[203] E. M. Voorhees and D. K. Harman. *TREC: Experiment and evaluation in information retrieval.* MIT Press, Cambridge, Massachusetts, USA, 2005.

[204] E. M. Voorhees and D. K. Harman. The Text REtrieval Conference. *TREC: Experiment and Evaluation in Information Retrieval*, pages 3–19, 2005. Chapter 1.

[205] E. M. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, and L. L. Wang. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1–12, 2020.

[206] J. Wang and J. Zhuhan. Portfolio theory of information retrieval. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 115–122, 2009.

[207] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 761–770, 2012.

[208] X. J. Wang, M. R. Grossman, and S. G. Hyun. Participation in TREC 2020 COVID track using continuous active learning. *arXiv preprint*, abs/2011.01453, 2020.

[209] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 571–580, 2008.

[210] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. on Information Systems*, 28(4):20, 2010.

[211] W. Webber, A. Moffat, and J. Zobel. The effect of pooling and evaluation depth on metric stability. In *Int. Workshop on Evaluating Information Access*, pages 7–15, 2010.

[212] W. E. Webber. *Measurement in information retrieval evaluation.* PhD thesis, The University of Melbourne, 2010. URL `http://hdl.handle.net/11343/35779`. (Accessed on 19th August 2022).

[213] E. W. Weisstein. “Erf” from *MathWorld*— A Wolfram web resource. `https://mathworld.wolfram.com/Erf.html`, 2022. (Accessed on 14th October 2022).

[214] E. W. Weisstein. “Bonferroni correction” from *MathWorld*— A Wolfram web resource. `https://mathworld.wolfram.com/BonferroniCorrection.html`, 2022. (Accessed on 9th October 2022).

[215] M. Wiboonrat. Risk anatomy of data center power distribution systems. In *Proc. IEEE Int. Conf. on Sustainable Energy Information Technology (SEIT)*, pages 674–679, 2008.

[216] P. Yang and J. Lin. Reproducing and generalizing semantic term matching in axiomatic information retrieval. In *Proc. European Conf. on Information Retrieval (ECIR)*, pages 369–381, 2019.

[217] J. Yi, Y. Chen, J. Li, S. Sett, and T. W. Yan. Predictive model performance: Offline and online evaluations. In *Proc. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 1294–1302, 2013.

[218] I. B. Yılmazel and A. Arslan. An intrinsic evaluation of the Waterloo spam rankings of the ClueWeb09 and ClueWeb12 datasets. *J. of Information Science*, 47(1):41–57, 2021.

[219] F. Zampieri, K. Roitero, J. S. Culpepper, O. Kurland, and S. Mizzaro. On topic difficulty in IR evaluation: The effect of systems, corpora, and system components. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 909–912, 2019.

[220] D. Zhang, J. Wang, E. Yilmaz, X. Wang, and Y. Zhou. Bayesian performance comparison of text classifiers. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 15–24, 2016.

[221] E. Zhang, N. Gupta, R. Tang, X. Han, R. Pradeep, K. Lu, Y. Zhang, R. Nogueira, K. Cho, H. Fang, and J. Lin. Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. *arXiv preprint*, abs/2007.07846, 2020.

[222] D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 17–20, 2010.

[223] J. Zhu, J. Wang, J. I. Cox, and M. J. Taylor. Risky business: Modeling and exploiting uncertainty in information retrieval. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 99–106, 2009.

[224] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 307–314, 1998.

Publications

Parts of this thesis previously appeared in the following refereed conferences:

- R. Benham, A. Moffat, and J. S. Culpepper. Bayesian System Inference On Shallow Pools. In *Proceedings of the 43rd European Conference On Information Retrieval (ECIR 2021).*
- R. Benham, B. Carterette, J. S. Culpepper, and A. Moffat. Bayesian Inferential Risk Evaluation On Multiple IR Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020).*
- R. Benham, B. Carterette, A. Moffat, and J. S. Culpepper. Taking Risks with Confidence. In *Proceedings of the 24th Australasian Document Computing Symposium (ADCS 2019).*
- R. Benham, A. Moffat, and J. S. Culpepper. On The Pluses and Minuses of Risk. In *Proceedings of the 15th Asia Information Retrieval Societies Conference (AIRS 2019).*