Statistical Significance Testing in Information Retrieval: Theory and Practice

Ben Carterette
University of Delaware
carteret@udel.edu

ABSTRACT

The past 20 years have seen a great improvement in the rigor of information retrieval experimentation, due primarily to two factors: high-quality, public, portable test collections such as those produced by TREC (the Text REtrieval Conference [38]), and the increased practice of statistical hypothesis testing to determine whether measured improvements can be ascribed to something other than random chance. Together these create a very useful standard for reviewers, program committees, and journal editors; work in information retrieval (IR) increasingly cannot be published unless it has been evaluated using a well-constructed test collection and shown to produce a statistically significant improvement over a good baseline.

But, as the saying goes, any tool sharp enough to be useful is also sharp enough to be dangerous. Statistical tests of significance are widely misunderstood. Most researchers and developers treat them as a "black box": evaluation results go in and a p-value comes out. But because significance is such an important factor in determining what research directions to explore and what is published, using p-values obtained without thought can have consequences for everyone doing research in IR. Ioannidis has argued that the main consequence in the biomedical sciences is that most published research findings are false [20]; could that be the case in IR as well?

CCS CONCEPTS

•Information systems →Evaluation of retrieval results; Presentation of retrieval results;

KEYWORDS

information retrieval; evaluation; statistical significance testing; reproducibility

1 OVERVIEW & OBJECTIVES

The past 20 years have seen a great improvement in the rigor of information retrieval experimentation, due primarily to two factors: high-quality, public, portable test collections such as those produced by TREC (the Text REtrieval Conference [38]), and the increased practice of statistical hypothesis testing to determine whether measured improvements can be ascribed to something other than random chance. Together these create a very useful standard for reviewers, program committees, and journal editors; work in information retrieval (IR) increasingly cannot be published unless it has been evaluated using a well-constructed test collection and shown to produce a statistically significant improvement over a good baseline.

But, as the saying goes, any tool sharp enough to be useful is also sharp enough to be dangerous. Statistical tests of significance are widely misunderstood. Most researchers and developers treat them as a "black box": evaluation results go in and a p-value comes out. But because significance is such an important factor in determining what research directions to explore and what is published, using p-values obtained without thought can have consequences for everyone doing research in IR. Ioannidis has argued that the main consequence in the biomedical sciences is that most published research findings are false [20]; could that be the case in IR as well?

This tutorial will help researchers and developers gain a better understanding of how tests work and how they should be interpreted so that they can both use them more effectively in their day-to-day work as well as better understand how to interpret them when reading the work of others. It will be appropriate for researchers and practitioners who are new to IR and wish to learn the standards of the community for significance testing and reporting, but also for experienced IR researchers and practitioners who already perform tests but desire a deeper understanding of what they are and how to interpret the information they provide. We aim to bridge the gap between the statistical theories of testing and how they are actually used in IR—a gap that is larger than is commonly understood.

The tutorial differs from previous iterations by the same presenter [7, 8, 10] by presenting expanded material on problems caused by sequential testing, multiple testing, and statistical issues related to generalizability of published results, while reducing material on the basics of statistical hypothesis testing for IR.

2 PROPOSER BIOGRAPHY

Ben Carterette is an Associate Professor of Computer and Information Sciences at the University of Delaware in Newark, Delaware, USA. His research primarily focuses on evaluation in Information Retrieval, including test collection construction, evaluation measures, and statistical testing. He has published over 80 papers in venues such as ACM TOIS, SIGIR, CIKM, WSDM, ECIR, and ICTIR, winning four Best Paper Awards for his work on evaluation. In addition, he has co-organized four workshops on IR evaluation and co-coordinated many TREC tracks. In addition to the ICTIR tutorials mentioned above, he has presented three SIGIR tutorials (one on low-cost evaluation, one on evaluation measures, and one

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: 10.1145/3077136.3080738

on statistical testing in IR) and one summer course for RUSSIR on evaluation in general.

Contact information: Ben Carterette, 101 Smith Hall, University of Delaware, Newark, DE, USA 19716. Phone: +1 302-831-3185. E-mail: carteret@cis.udel.edu. E-mail preferred.

3 OUTLINE

The tutorial will be organized and presented in three parts. The first part will start with a brief introduction to the reasons for testing significance and a guide to the t-test and empirical tests (the randomization and bootstrap tests), which provide a good introduction to the underlying ideas necessary for a Bayesian framework for testing. We will develop a Bayesian version of the t-test and show how it can be extended easily. The second part will dig into the theory of testing, with the goal of providing a deeper understanding about exactly what significance tests tell us, what their limitations are, and how misunderstanding them can lead us astray. A large part of the focus of this section will be on problems that arise from multiple testing, sequential testing, and the use of "extreme" values (that is, the highest observed value of some effectiveness measure). The third part will take a broader view, looking at how significance tests are used by scientists & engineers in the field and by the reviewers and editors of conferences and journals. In this part we will consider the interpretation of significance tests in terms of reproducibility and generalizability.

(1) Testing significance in IR (35 min)
  (a) Why test significance? (5 min)
  (b) Common tests used in IR (10 min)
    (i) t-test (including ANOVA and the linear model)
    (ii) randomization and bootstrap tests
  (c) Bayesian framework (20 min)
(2) Theory of significance testing (and what it means for you) (95 min)
  (a) Terms and definitions—a test-independent foundation (25 min)
    (i) null hypothesis and alternate hypothesis
    (ii) paired (one-sample) and unpaired (two-sample); one-tailed and two-tailed
    (iii) test statistic; confidence interval; p-value; critical value; effect size
    (iv) power and accuracy; false positives and false negatives in testing
    (v) sources of variance; within-group and between-group variance
  (b) Myths and misconceptions (30 min)
    (i) statistical significance has an intrinsic meaning
    (ii) when data violates the assumptions of a test, the test cannot be used
    (iii) the p-value has an intrinsic meaning
    (iv) smaller p-values indicate greater significance
    (v) a p-value less than 0.05 indicates "significance"
  (c) Multiple testing, sequential testing, and extreme values (40 min)
    (i) testing many hypotheses simultaneously
    (ii) testing the same hypothesis again and again
    (iii) re-using test collections to test the same hypotheses
    (iv) interpreting the highest measured effectiveness from a series of tests
(3) How significance testing affects us all (50 min)
  (a) Three classes of scientist engineers and how they make use of significance tests: (10 min)
    (i) as readers of research papers: to guide choice of baseline systems
    (ii) as working scientists/engineers: to guide experimentation and determine what to publish or deploy
    (iii) as reviewers and editors: to guide publication recommendations and decisions
  (b) Reproducibility and generalizability (20 min)
    (i) what does significance tell us about reproducibility?
    (ii) what does significance tell us about generalizability?
    (iii) how to assess generalizability?
  (c) How adherence to the misconceptions listed above can set back research and development in IR (10 min)
  (d) Specific recommendations for using and interpreting tests (10 min)

4 MATERIALS

Tutorial materials, including presentation slides, R labs, and worksheets, will be available on the web at http://ir.cis.udel.edu/SIGIR17tutorial.

5 ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation (NSF) under grant number IIS-1350799. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Jaime Arguello, Matt Crane, Fernando Diaz, Jimmy Lin, and Andrew Trotman. Report on the sigir 2015 workshop on reproducibility, inexplicability, nd generalizability of results (rigor). *SIGIR Forum*, 49(2), 2015.
[2] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of CIKM*, pages 601–610, 2009.
[3] James O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32.
[4] Leonid Boytsov, Anna Belova, and Peter Westfall. Deciding on an adjustment for multiplicity in IR experiments. In *Proceedings of SIGIR*, 2013.
[5] Ben Carterette. Model-based inference about ir systems. In *Proceedings of ICTIR*, 2011.
[6] Ben Carterette. Multiple testing in statistical analysis of systems-based IR experiments. *ACM TOIS*, 2012.
[7] Ben Carterette. Statistical significance testing in information retrieval: Theory and practice. In *Proceedings of ICTIR*, 2013.
[8] Ben Carterette. Statistical significance testing in information retrieval: Theory and practice. In *Proceedings of SIGIR*, 2014.
[9] Ben Carterette. Bayesian inference for information retrieval evaluation. In *Proceedings of ICTIR*, 2015.
[10] Ben Carterette. Statistical significance testing in information retrieval: Theory and practice. In *Proceedings of ICTIR*, 2015.
[11] Ben Carterette, Evangelos Kanoulas, Virgil Pavlu, and Hui Fang. Building reusable test collections through experimental design. In *Proceedings of SIGIR*, 2010.
[12] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for systems effectiveness evaluation. In *Proceedings of CIKM*, 2011.

[13] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proceedings of CIKM*, 2012.

[14] Ben Carterette and Mark D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*, pages 643–652, 2007.

[15] G. V. Cormack and T. R. Lyman. Statistical precision of information retrieval evaluation. In *Proceedings of SIGIR*, pages 533–540, 2006.

[16] Nicola Ferro. Reproducibility challenges in information retrieval. *Journal of Data and Information Quality*, 8, 2017.

[17] Nicola Ferro and Gianmaria Silvello. A general linear mixed models approach to study system component effects. In *Proceedings of SIGIR*, 2016.

[18] Alan Hanbury and Henning Müller. Automated component-level evaluation: Present and future. In *Proceedings of CLEF*, 2010.

[19] John P. A. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2):218–228, 2005.

[20] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005.

[21] Douglas H. Johnson. The insignificance of statistical significance testing. Technical Report 225, USGS Northern Prarie Wildlife Research Center, 1999.

[22] Karen Sparck Jones, editor. *Information Retrieval Experiment*. Buttersworth, 1981.

[23] Karen Sparck Jones and Peter Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, 1997.

[24] Alistair Moffat and Justin Zobel. What does it mean to "measure performance"? In *Proceedings of WISE*, pages 1–12, 2004.

[25] Regina Nuzzo. Statistical errors. *Nature News*, 506, 2014.

[26] Tetsuya Sakai. Topic set size design and power analysis in practice. In *Proceedings of ICTIR*, 2016.

[27] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[28] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of SIGIR*, pages 162–169, 2005.

[29] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM*, pages 623–632, 2007.

[30] Mark D. Smucker, James Allan, and Ben Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of SIGIR*, pages 630–631, 2009.

[31] Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P. A. Ioannidis, and Michela Taufer. Enhancing reproducibility for computational methods. *Science*, 354, 016.

[32] Jean Tague. The pragmatics of information retrieval evaluation. In Jones [22], pages 59–102.

[33] Jean Tague-Sutcliffe. The pragmatics of information retrieval evaluation revisited. In Jones and Willett [23], pages 205–216.

[34] Jean Tague-Sutcliffe and James Blustein. A statistical analysis of the TREC-3 data. In *Proceedings of the 3rd Text REtrieval Conference (TREC)*, pages 385–399, 1994.

[35] Julián Urbano, Mónica Marrero, and Diego Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proceedings of SIGIR*, 2013.

[36] Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR*, pages 315–323, 1998.

[37] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of SIGIR*, pages 316–323, 2002.

[38] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiments and evaluation in information retrieval*. The MIT Press, 2005.

[39] William Webber, Alistair Moffat, and Justin Zobel. Statistical power in retrieval experimentation. In *Proceedings of CIKM*, pages 571–580, 2008.

[40] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.