Statistical Analysis of Results in Music Information Retrieval: Why and How

A Tutorial at ISMIR 2018

Julián Urbano
Delft University of Technology
The Netherlands

Arthur Flexer
Austrian Research Institute for Artificial Intelligence
Austria

1. OVERVIEW AND OBJECTIVES

Nearly since the beginning, the ISMIR and MIREX communities have promoted rigor in experimentation through the creation of datasets and the practice of statistical hypothesis testing to determine the reliability of the improvements observed with those datasets. In fact, MIR researchers have adopted a certain way of going about statistical testing, namely non-parametric approaches like the Friedman test and multiple comparisons corrections like Tukey's. In a way, they have become a standard for researchers reporting their results, reviewers, committees, journal editors, etc. It is nowadays more frequent to require statistically significant improvements over a baseline with a well-established dataset.

But hypothesis testing can be very misleading if not well understood. To many researchers, especially newcomers, even the simpler analyses and tests are seen as a black box where one puts performance scores and gets a p-value which, as they are told, must be smaller than 0.05. Therefore, significance tests are in part responsible of determining what gets published, what research lines to follow, and what project to fund, so it is very important to understand what they really mean and how they should be carried out and interpreted.

The goal of this tutorial is to help MIR researchers and developers get a better understanding of how these statistical methods work and how they should be interpreted. Starting from the very beginning of the evaluation process, it will show that statistical analysis is always required, but that too much focus on it, or the incorrect approach, is just harmful. The tutorial will attempt to provide better insight into statistical analysis of results, present better solutions and guidelines, and point the attendees to the larger but ignored problems of evaluation and reproducibility in MIR.

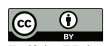
The slides of the tutorial are available online at `https://www.slideshare.net/caerolus/statistical-analysis-of-results-in-music-information-retrieval-why-and-how`.

2. OUTLINE OF THE TUTORIAL

The tutorial will be divided in four main blocks:

I. Why? Evaluation, hypotheses and experiments (30 mins). This block will review the typical evaluation process from the basics, starting from the task definition to reporting the results in a paper. Even if short, a full review is necessary to cover what the problem really is, why we do things the way we do them, what we assume in the process, and what it is that we report in reality. In particular, we will show that the evaluation problem is basically an estimation problem that goes through the steps of sampling, measurement, estimation, inference and generalization. From here we will make clear that evaluation experiments are subject to systematic and random error, so we need to carry out statistical analyses and worry about validity and reliability. This distinction is important because any analysis assumes validity and focuses on reliability. Later on the tutorial we will see how validity is neglected even though it is the biggest of our problems.

II. How? Hypothesis testing, details and misconceptions (70 mins). Here we will cover the basics of statistical hypothesis testing and how it has evolved from the Fisher and Neyman-Pearson frameworks until what we do nowadays. A historical perspective is a great way to make the audience understand the inner workings of hypothesis testing and what we try to achieve with it. We will cover typical tests and common variations, which will lead us to the fact


© Julián Urbano, Arthur Flexer. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Julián Urbano, Arthur Flexer. "Statistical Analysis of Results in Music Information Retrieval: Why and How", Tutorial presented at the 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

that our results and conclusions are based on models that make blunt assumptions. The common details of current statistical hypothesis testing, such as accuracy, power and error types, will be discussed next, leading to the problem of multiple comparisons and how it is typically tackled in MIR research with the Friedman-Tukey procedure. This is the perfect spot to make the audience aware of some typical myths and misconceptions around hypothesis testing.

III. **What else? Validity of MIR experiments (60 mins).** This block will focus on neglected validity, i.e. whether an experiment really measures what one wants to examine. We will show how a lack of internal or external validity, even if experiments are reliable and repeatable and hypothesis testing is done correctly, can severely impede progress in MIR. One example are adversarial examples, which are marginally and imperceptibly altered data that are able to drastically reduce performance of MIR systems. It has even been claimed that such easily fooled MIR systems do not use musical knowledge at all. Another example we will discuss is the lack of inter-rater agreement when annotating ground truth data, which is a validity problem responsible for so-called glass ceilings, i.e. that performance in many MIR tasks can never exceed a certain upper level.

IV. **So? Guidelines and reform (20 mins).** This last block will wrap up the tutorial with a summary of main concepts and their interplay, such as p-values, falseness of null hypotheses and their use in decision-making, assumptions, replication and relevance of research. Guides will be provided for better statistical analysis in MIR, and arguments will be made for a change in the statistical practice in the field. The four main takeaways are:

- There is always random error in our experiments, so we always need some kind of statistical analysis.
- There is no point in being too picky or intense about how we do it.
- Nobody knows how to do it properly, and different fields adopt different methods.
- What is far more productive, is to adopt an exploratory attitude rather than mechanically testing.

3. INTENDED AUDIENCE

The tutorial is aimed at people covering various roles in MIR research. It will be interesting to novice researchers who probably see the evaluation process as a black box from which they obtain a p-value which is hopefully below 0.05. More experienced researchers should be interested in the tutorial to strengthen their understanding of statistical analysis and make better decisions during their research and when evaluating someone else's work for publication or funding applications. All attendees, researchers or not, could benefit from the tutorial because of their role as readers and consumers of MIR literature. Previous knowledge of the basics of the evaluation process and of probability theory are desirable but not required.

4. PRESENTERS

Julián Urbano is an Assistant Professor at Delft University of Technology, The Netherlands. His research is primarily concerned with evaluation in IR, working in both the music and text domains. Current topics of interest are the application of statistical methods for the construction of datasets, the reliability of evaluation experiments, statistical significance testing for IR, low-cost evaluation and stochastic simulations for evaluation. He has published over 50 research papers in related venues like Foundations and Trends in IR, the IR Journal, the Journal of Multimedia IR, ISMIR, CMMR, ACM SIGIR, ACM CIKM and ECIR, winning two best paper awards and an outstanding reviewer award. He has been active in the ISMIR community since 2010, both as author and PC member. He is reviewer for other conferences and journals, such as ACM CIKM, HCOMP, IEEE TASLP, IEEE MM, ACM TWEB, IEEE TKDE or the Information Sciences journal.

Julián has also extensive experience in teaching both at the graduate and undergraduate levels. He has accumulated over 600 classroom hours, has received several awards for his teaching activities, and has published two papers related to Computer Science education.

Arthur Flexer is a senior researcher, project manager and vice-head at 'The Intelligent Music Processing and Machine Learning Group' of the Austrian Research Institute for Artificial Intelligence (OFAI). He has twenty-five years of experience in basic research on machine learning with an emphasis on applications to music in the last twelve years. He holds a PhD in psychology which provides him with the necessary background concerning design and evaluation of experiments. He has published comprehensively on the role of experiments and on problems of ground truth and inter-rater agreement, all in the field of MIR. He is author and co-author of more than 80 peer-reviewed articles. He has been active in the ISMIR community since 2005 and has also published in related venues like DAFx, SMC, ECIR, Journal of Machine Learning Research and Journal of New Music Research. He is a member of the editorial board of the Transactions of the International Society for Music Information Retrieval (TISMIR).

During his time as an Assistant Professor at the Institute of Medical Cybernetics and Artificial Intelligence (Center for Brain Research, Medical University Vienna, 2005-2007) he has been teaching courses on Machine Learning, Data Mining and Statistical Evaluation.

5. ACKNOWLEDGEMENTS

J. Urbano was supported by the European Commission H2020 project TROMPA (770376-2). A. Flexer was supported by the Vienna Science and Technology Fund (WWTF, project MA14-018).

6. REFERENCES

[1] Al-Maskari, A., Sanderson, M., & Clough, P. (2007). The Relationship between IR Effectiveness Measures and User Satisfaction. ACM SIGIR
[2] Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. Journal of Wildfire Management
[3] Armstrong, T.G., Moffat, A., Webber, W. & Zobel, J. (2009). Improvements that don't add up: ad-hoc retrieval results since 1998. CIKM
[4] Balke, S., Driedger, J., Abeßer, J., Dittmar, C. & Müller, M. (2016). Towards Evaluating Multiple Predominant Melody Annotations in Jazz Recordings. ISMIR
[5] Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? Statistical Science
[6] Bosch J.J. & Gómez E. (2014). Melody extraction in symphonic classical music: acomparative study of mutual agreement between humans and algorithms. Conference on Interdisciplinary Musicology
[7] Boytsov, L., Belova, A. & Westfall, P. (2013). Deciding on an adjustment for multiplicity in IR experiments. SIGIR
[8] Carterette, B. (2012). Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. ACM Transactions on Information Systems
[9] Carterette, B. (2015a). Statistical Significance Testing in Information Retrieval: Theory and Practice. ICTIR
[10] Carterette, B. (2015b). Bayesian Inference for Information Retrieval Evaluation. ACM ICTIR
[11] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum
[12] Cormack, G. V., & Lynam, T. R. (2006). Statistical Precision of Information Retrieval Evaluation. ACM SIGIR
[13] Downie, J. S. (2004). The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future. Computer Music Journal
[14] Fisher, R. A. (1925). Statistical Methods for Research Workers. Cosmo Publications
[15] Flexer, A. (2006). Statistical Evaluation of Music Information Retrieval Experiments. Journal of New Music Research
[16] Flexer, A., Grill, T.: The Problem of Limited Inter-rater Agreement in Modelling Music Similarity, Journal of New Music Research
[17] Gelman, A. (2013b). The problem with p-values is how they're used.
[18] Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness
[19] Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no shing expedition' or p-hacking' and the research hypothesis was posited ahead of time.
[20] Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. American Scientist
[21] Gelman, A., & Stern, H. (2006). The Difference Between Significant and Not Significant is not Itself Statistically Significant. The American Statistician
[22] Goodfellow I.J., Shlens J. & Szegedy C. (2014). Explaining and harnessing adversarial examples. ICLR
[23] Gouyon, F., Sturm, B. L., Oliveira, J. L., Hespanhol, N., & Langlois, T. (2014). On Evaluation Validity in Music Autotagging. ACM Computing Research Repository.
[24] Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., & Olson, D. (2000). Do Batch and User Evaluations Give the Same Results? ACM SIGIR
[25] Hu, X., & Kando, N. (2012). User-Centered Measures vs. System Effectiveness in Finding Similar Songs. ISMIR
[26] Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. ACM SIGIR
[27] Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLoS Medicine
[28] Lehmann, E.L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Journal of the American Statistical Association
[29] Lehmann, E.L. (2011). Fisher, Neyman, and the Creation of Classical Statistics. Springer
[30] Lee, J. H., & Cunningham, S. J. (2013). Toward an understanding of the history and impact of user studies in music information retrieval. Journal of Intelligent Information Systems
[31] Marques, G., Domingues, M. A., Langlois, T., & Gouyon, F. (2011). Three Current Issues In Music Autotagging. ISMIR
[32] Neyman, J. & Pearson, E.S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. Biometrika
[33] Panteli, M., Rocha, B., Bogaards, N. & Honingh, A. (2017). A model for rhythm and timbre similarity in electronic dance music. Musicae Scientiae
[34] Quinton, E., Harte, C. & Sandler, M. (2015). Extraction of metrical structure from music recordings. DAFX
[35] Sakai, T. (2014). Statistical Reform in Information Retrieval? ACM SIGIR Forum
[36] Savoy, J. (1997). Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing and Management
[37] Schedl, M., Flexer, A., & Urbano, J. (2013). The Neglected User in Music Information Retrieval Research. Journal of Intelligent Information Systems
[38] Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin
[39] Serrà, J., Müller, M., Grosche, P., & Arcos, J.L. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. IEEE Trans. on Multimedia
[40] Smith, J.B.L. & Chew, E. (2013). A meta-analysis of the MIREX structure segmentation task. ISMIR
[41] Smucker, M. D., Allan, J., & Carterette, B. (2007). A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. ACM CIKM
[42] Smucker, M. D., Allan, J., & Carterette, B. (2009). Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes. CM SIGIR
[43] Smucker, M. D., & Clarke, C. L. A. (2012). The Fault, Dear Researchers, is Not in Cranfield, But in Our Metrics, that They Are Unrealistic. European Workshop on Human-Computer Interaction and Information Retrieval
[44] Student. (1908). The Probable Error of a Mean. Biometrika
[45] Sturm, B. L. (2013). Classification Accuracy is Not Enough: On the Evaluation ofMusic Genre Recognition Systems. Journal of Intelligent Information Systems
[46] Sturm, B. L. (2014). The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. Journal of New Music Research
[47] Sturm, B.L. (2014). A simple method to determine if a music information retrieval system is a horse, IEEE Trans. on Multimedia
[48] Sturm B.L. (2016). "The Horse" Inside: Seeking Causes Behind the Behaviors of Music Content Analysis Systems, Computers in Entertainment
[49] Tague-Sutcliffe, J. (1992). The Pragmatics of Information Retrieval Experimentation, Revisited. Information Processing and Management
[50] Turpin, A., & Hersh, W. (2001). Why Batch and User Evaluations Do Not Give the Same Results. ACM SIGIR

[51] Urbano, J. (2015). Test Collection Reliability: A Study of Bias and Robustness to Statistical Assumptions via Stochastic Simulation. Information Retrieval Journal
[52] Urbano, J., Downie, J. S., McFee, B., & Schedl, M. (2012). How Significant is Statistically Significant? The case of Audio Music Similarity and Retrieval. ISMIR
[53] Urbano, J., Marrero, M., & Martín, D. (2013a). A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation. ACM SIGIR
[54] Urbano, J., Marrero, M. & Martín, D. (2013b). On the Measurement of Test Collection Reliability. SIGIR
[55] Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in Music Information Retrieval. Journal of Intelligent Information Systems
[56] Urbano, J. & Marrero, M. (2016). Toward Estimating the Rank Correlation between the Test Collection Results and the True System Performance. SIGIR
[57] Urbano, J. & Nagler, T. (2018). Stochastic Simulation of Test Collections: Evaluation Scores. SIGIR
[58] Voorhees, E. M., & Buckley, C. (2002). The Effect of Topic Set Size on Retrieval Experiment Error. ACM SIGIR
[59] Webber, W., Moffat, A., & Zobel, J. (2008). Statistical Power in Retrieval Experimentation. ACM CIKM
[60] Ziliak, S. T., & McCloskey, D. N. (2008). The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press
[61] Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments? ACM SIGIR