Evaluating Search-Based Techniques With Statistical Tests

Andre Arcuri
Westerdals Oslo ACT, Faculty of Technology, Oslo, Norway,
and SnT, University of Luxembourg, Luxembourg.
arcand@westerdals.no

ABSTRACT

This tutorial covers the basics of how to use statistical tests to evaluate and compare search-algorithms, in particular when applied on software engineering problems. Search-algorithms like Hill Climbing and Genetic Algorithms are randomised. Running such randomised algorithms twice on the same problem can give different results. It is hence important to run such algorithms multiple times to collect average results, and avoid so publishing wrong conclusions that were based on just luck. However, there is the question of how often such runs should be repeated. Given a set of n repeated experiments, is such n large enough to draw sound conclusions? Or should had more experiments been run? Statistical tests like the Wilcoxon-Mann-Whitney U-test can be used to answer these important questions.

CCS CONCEPTS

• **Software and its engineering** → *Software testing and debugging*; *Search-based software engineering*;

KEYWORDS

Statistics, Tutorial, Non-parametric Test, Effect Size, SBSE, SBST

ACM Reference format:
Andre Arcuri. 2018. Evaluating Search-Based Techniques With Statistical Tests. In *Proceedings of SBST'18:IEEE/ACM 11th International Workshop on Search-Based Software Testing , Gothenburg, Sweden, May 28--29, 2018 (SBST'18),* 1 pages.
https://doi.org/10.1145/3194718.3194732

TUTORIAL DESCRIPTION

This tutorial is based on the guidelines on statistical tests published in [1, 2]. Further information can be found for example in [3--8].

The tutorial covers these topics:

- Motivation for using statistical tests.
- Pitfalls when comparing search algorithms.
- Statistical difference and p-values.
- Fisher Exact test.
- Wilcoxon-Mann-Whitney U-test.
- Parametric vs. non-parametric tests.
- Standardised Effect Sizes like Vargha-Delaney $\hat{A}_{12}$.
- Multiple experiments/comparisons.
- Result visualisation in Latex/R/Python.

The main takeaways from this tutorial are:

- Statistics is very important and used in all fields of science and engineering. When dealing with empirical experiments in which search algorithms are involved, the use of statistics should be considered mandatory.
- Statistics is a complex topic. When a researcher is not sure on which tests to use, non-parametric ones should be preferred. However, the more data is collected, the less important the statistical tests become.
- The 0.05 p-value threshold for statistical significance is *arbitrary*. When running experiments with large clusters of computers, it can become trivial to achieve statistical significant results on any experiment. One should always also look if the results are of practical importance. Effect sizes like the Vargha-Delaney $\hat{A}_{12}$ should be used to help in this regard.
- The data analysis should be automated with scripts written in languages like R and Python. As experiments can and will be repeated (e.g., when a bug is found, or new artefacts are added to the case study), it is important that the papers can be rebuilt (e.g., creation of PDF files after updating all the tables and figures) automatically. This is rather straightforward to do when using professional tools like Latex. Not only it would result in a massive amount of saved time, but also it reduces the risk of introducing mistakes when editing tables manually right just before a paper submission deadline.

ACKNOWLEDGMENTS

This work is supported by the National Research Fund, Luxembourg (FNR/P10/03).

REFERENCES

[1] A. Arcuri and L. Briand. 2011. A Practical Guide for Using Statistical Tests to Assess Randomized Algorithms in Software Engineering. In *ACM/IEEE International Conference on Software Engineering (ICSE)*. 1--10.

[2] A. Arcuri and L. Briand. 2014. A Hitchhiker's Guide to Statistical Tests for Assessing Randomized Algorithms in Software Engineering. *Software Testing, Verification and Reliability (STVR)* 24, 3 (2014), 219--250.

[3] M. Cowles and C. Davis. 1982. On the origins of the .05 level of statistical significance. *American Psychologist* 37, 5 (1982), 553--558.

[4] M.P. Fay and M.A. Proschan. 2010. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys* 4 (2010), 1--39.

[5] J.P.A. Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.

[6] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50--60.

[7] T.V. Perneger. 1998. What's wrong with Bonferroni adjustments. *British Medical Journal* 316 (1998), 1236--1238.

[8] A. Vargha and H. D. Delaney. 2000. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 25, 2 (2000), 101--132.

SBST'18, May 28--29, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5741-8/18/05.
https://doi.org/10.1145/3194718.3194732