

Illustration of Statistical Test Results for Experiment Evaluation

Thomas Weise

Nature Inspired Computation and Applications Laboratory (NICAL),
School of Computer Science and Technology,
University of Science and Technology of China;
Hefei, Anhui, China, 230027.
<http://nical.ustc.edu.cn> · <http://www.it-weise.de>
tweise@ustc.edu.cn · tweise@gmx.de

Abstract In this reference document, I provide some notes on how I (personally) think that the results of statistical tests involving $N > 2$ processes could be represented in a compact and easy-to-read way. The proposed method is a directed graph acyclic graph (DAG), which is based on the fact that the outcomes of tests always are at least a strict partial order which can be visualized as DAG.

Version: August 24, 2011

1 Introduction

In many research areas, experiments are performed in order to find which method is best to solve a given problem. In the field of numerical optimization methods, for example, we often apply different metaheuristic algorithms [1] a couple of times to the same set of benchmark functions. Then, statistical tests such as the Mann-Whitney U test Mann and Whitney [2], Siegel and Castellan Jr. [3], Dinneen and Blakesley [4], Neumann [5] are used to compare the results achieved with the different methods. The outcome of such a test, when comparing two such datasets, is usually something like *With a probability to err of no more than 2% (i. e., at a significance level of 2%), we can state that “Method A” outperforms “Method B”* or *At a significance level of 5% (or with a maximally allowed error probability of 5%), no statistically significant difference can be detected between the performances of “Method A” and “Method B”*.

In this reference document we want to introduce a very simple way to visualize the outcomes of statistical tests used for comparing N processes (or distributions) based on datasets sampled from them. We have used this simple approach in the past in some of our research works (such as [6–8]) and have been both, praised and criticized by reviewers for its use. The positive feedback mainly was based on the simplicity and clarity of the presentation, the negative feedback contained complains about the non-standard way of representing

such data. With this work we thus intend to somewhat standardize a simple and compact graphical notation for results of statistical tests.

Generally, statistical tests [3, 9–13] are tools to compare processes that produce measurable outputs which can be represented as real numbers. Often, two such processes P_1 and P_2 are compared with the goal to find which of the two tends to produce smaller (or larger) outputs. Given finite samples (observations) of these processes, this question can be answered by applying statistical tests such as, for example, the Mann-Whitney U test. Based on a significance level α , i.e., a threshold for the highest acceptable probability to make a false statement, a significant difference between P_1 and P_2 is either confirmed or rejected.

2 Illustration of $N(N - 1)/2$ Comparisons

If $N > 2$ processes P_1, P_2, \dots, P_N are observed, the previous question can be extended to finding which of them tends to contain the smallest elements and to detect interrelations. One way to do this is to compare each process with every other process, again using the statistical test of choice. This will result in $N(N - 1)/2$ outcomes. A common way to represent these outcomes is to use a table (matrix) $T_{i,j} \in \{+, -, 0\}$. A value of $T_{i,j} = +$ in the i^{th} row and j^{column} means that process P_i has significantly larger outputs than process P_j , a $-$ stands for smaller outputs, and 0 symbolizes that no significant difference could be detected (at the given significance level α).

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}
P_1		+	+	+	+	+	0	+	+	+	+
P_2			0	+	0	0	-	+	0	-	-
P_3				+	-	0	-	0	0	0	0
P_4					-	-	-	-	-	-	-
P_5						0	-	0	0	0	0
P_6							-	+	0	-	-
P_7								+	+	+	+
P_8									-	-	-
P_9										-	-
P_{10}											0
P_{11}											

Table 1: An example for a table specifying the outcome of the statistical comparison of eleven processes P_1 to P_{11} .

Table 1 gives an example how a common tabular illustration of the comparison results for eleven processes P_1 to P_{11} could look like. Only the upper triangle of the table needs to be populated since $T_{i,j} = + \Rightarrow T_{j,i} = -$, $T_{i,j} = - \Rightarrow T_{j,i} = +$, $T_{i,j} = 0 \Rightarrow T_{j,i} = 0$, and $T_{i,i} = 0$ for all $i, j \in 1..N$. From the example, it becomes clear that with a rising number of processes, it becomes quite hard to recognize the order of the processes according to the tests from such a table.

3 Graph-based Notation

3.1 Example

Indeed, a full set of $N(N - 1)/2$ test results defines a (strict) partial order on the compared processes. Besides using a table or matrix, such a partial order can be illustrated as directed acyclic graph (DAG), as sketched in Fig. 1.a. Each process can be represented as a node in a graph. Here, $T_{i,j} = +$ will result in a directed edge from the node labeled with P_j to the node labeled with P_i . A $-$ results in a directed edge into the opposite direction and a 0 is represented by adding no edge between the corresponding nodes.

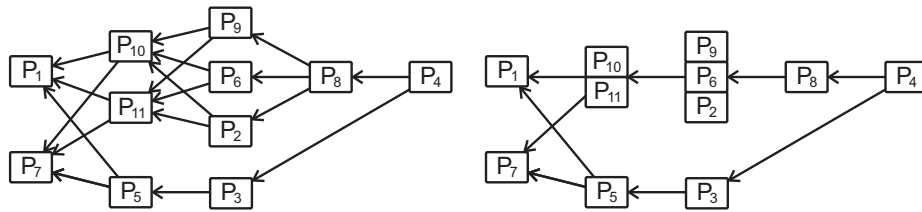


Fig. 1.a: full graph

Fig. 1.b: simplified graph

Figure 1: The example results from Table 1 illustrated as graph.

Since the test results form a transitive order, edges which are sufficiently explained by transitivity can be omitted in the graph (and actually, the corresponding tests did not need to be performed in the first place). Hence Fig. 1.b does not contain an arrow from node P_2 to P_1 , since that one is already subsumed by the arrow from P_2 to P_{11} and from P_{11} to P_1 .

The graph sketched in Fig. 1.a is faster to read than Table 1. The notion can further be simplified by combining those nodes for which all incoming arrows come from the same origins and all outgoing arrows target the same nodes. Fig. 1.b represents such a simplification. It is our strong belief that this representation should be preferred to a tabular representation, because of its compactness, clarity, and ease of use.

From Fig. 1.b, it can immediately be seen that processes P_1 and P_7 tend to have the largest outputs while P_4 has the smallest. There is no significant difference between P_9 and P_6 or P_3 , but P_9 tends to produce smaller outputs than P_{10} . The outputs of P_5 tend to be larger than those of P_3 , but they are no significantly different from those of P_2 .

3.2 Formal Definition

Given a set \mathbb{P} of N processes $P_i : i \in 1..N$ and a statistical test result matrix $T_{i,j} \in \{+, -, 0\} \forall i, j \in 1..N$, the graph-based representation G is defined as follows:

1. For each $P_i \in \mathbb{P}$, there exists exactly one node labeled with P_i in G .
2. A node may be labeled with a set S of multiple process names if and only if $\forall P_i, P_j \in S \Rightarrow (\forall P_k \in \mathbb{P} : (k \neq i) \wedge (k \neq j) \Rightarrow T_{i,k} = T_{j,k})$ holds.
3. There exists a directed edge from the node labeled with P_j to the node labeled P_i if and only if:
 - (a) $T_{i,j} = +$ (and, hence, $T_{j,i} = -$) and
 - (b) $\neg \exists P_k \in \mathbb{P} : T_{i,k} = + \wedge T_{k,j} = +$.

3.3 How to Use

The author wants to emphasize that a diagram such as Fig. 1.b should always be accompanied by a textual note stating the applied test and the test's configuration, the significance level, and the meaning of the presence of a directed edge in the graph. An example for this notion could be: *Fig. 1.b* shows the outcome of the application of a two-tailed Mann-Whitney U test with a significance level of 2% to the data sampled from P_1 to P_{11} . A directed edge from a node P_i to a node P_j means that, according to the applied test, P_i produces { larger / smaller / better } outcomes than P_j .

4 Afterwords

4.1 Recommendation for Experimenting in the Field of Optimization

For comparing the results different optimization algorithms [1], I personally recommend using the Mann-Whitney U test [2, 3]. This test has two advantages:

1. it is non-parametric, i. e., does not make the assumption that the results are normally distributed and
2. the compared datasets can contain different numbers of samples.

The first advantage may seem puzzling. Many statistical tests make the assumption that the datasets contain samples which, approximately, are normally distributed around some mean value (which also is the median). For sufficiently large datasets, we often assume that this assumption holds. However, at least in my opinion, when comparing the results of optimization runs, this is a dangerous idea:

1. Each sample in this case is the result of an optimization process and
2. the result of an optimization process, in this case, is the objective value of the best solution discovered during its course.
3. Objective functions are usually bounded, i. e., there exists at least one global optimum (and also at least one globally worst solution).
4. The normal distribution, on the other hand, is unbounded to all sides.
5. Optimization tends to find solutions closer to the optimum, so the median of the distribution of values in the datasets are not only closer to one of the bounds, but the distribution could also be skew.
6. Furthermore, there may exist gaps in the objective value spectrum, i. e., function values which cannot be reached.

Of course, for sufficiently many samples, the normal distribution assumption can be accepted. But to be on the safe side, I would always pick a non-parametric test.

4.2 Utilities

Mann-Whitney U Tester Along with this document, I provide a small utility which can construct tables similar to Table 1 based on the Mann-Whitney U test [2, 3] by parsing a directory of .txt files and comparing their contents. These .txt files contain one (textually-represented) real number per line. The resulting table will be stored in a file called results.txt.

```

1 E:\>java -jar mannWhitneyUTest.jar -dir:e:\x
2 Mann-Whitney U Tester
3   by Thomas Weise, tweise@gmx.de, http://www.it-weise.de/
4   statistically compares all .txt files in a directory
5   =====
6 -better:{smaller,bigger}
7   -> bigger values win
8 -level:XXX -> significance level
9   -> significance level 0.02
10 -tails:{one,two}
11   -> two tailed tests
12 -dir:path
13   -> dir E:\x
14
15
16 ===== Loading data =====
17
18 - loading E:\x\A.txt
19 - loading E:\x\B.txt
20 - loading E:\x\C.txt
21

```

```
22
23 ===== Done loading data: 3 files collected =====
24
25 Writing output to E:\x\results.txt
26 ===== done =====
```

Listing 1.1: The utility – invocation.

```
1      A B C
2  A    0 0
3  B     -
4  C
```

Listing 1.2: The utility – output.

References

- [1] Thomas Weise. *Global Optimization Algorithms – Theory and Application*. it-weise.de (self-published): Germany, 2009. URL <http://www.it-weise.de/projects/book.pdf>.
- [2] Henry B. Mann and Donald R. Whitney. On a Test of whether One of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947. 10.1214/aoms/1177730491. URL <http://projecteuclid.org/euclid.aoms/1177730491>.
- [3] Sidney Siegel and N. John Castellan Jr. *Nonparametric Statistics for The Behavioral Sciences*. Humanities/Social Sciences/Languages. McGraw-Hill: New York, NY, USA, 1956. ISBN 0-07-057357-3 and 070573434X.
- [4] L. C. Dinneen and B. C. Blakesley. Algorithm AS 62: A Generator for the Sampling Distribution of the Mann-Whitney U Statistic. *Journal of the Royal Statistical Society: Series C – Applied Statistics*, 22(2):269–273, 1973.
- [5] N. Neumann. Some Procedures for Calculating the Distributions of Elementary Non-parametric Test Statistics. *Statistical Software Newsletter (SSN)*, 14(3), 1988.
- [6] Thomas Weise. *Evolving Distributed Algorithms with Genetic Programming*. PhD thesis, University of Kassel, Fachbereich 16: Elektrotechnik/Informatik, Distributed Systems Group: Kassel, Hesse, Germany, May 4, 2009. URL <http://www.it-weise.de/documents/files/w2009DISS.pdf>. Won the Dissertation Award of The Association of German Engineers (Verein Deutscher Ingenieure, VDI).
- [7] Thomas Weise and Ke Tang. Evolving Distributed Algorithms with Genetic Programming. *IEEE Transactions on Evolutionary Computation (IEEE-EC)*, to appear, 2011.
- [8] Mingxu Wan, Thomas Weise, and Ke Tang. Novel Loop Structures and the Evolution of Mathematical Algorithms. In James A. Foster and Sara Silva, editors, *Proceedings of the 14th European Conference on Genetic Programming (EuroGP’11)*, Lecture Notes in Computer Science (LNCS). Springer-Verlag GmbH: Berlin, Germany, 2011. Nominated for best paper.
- [9] Lorenz Gygax. Statistik für Nutztierethologen – Einführung in die statistische Denkweise: Was ist, was macht ein statistischer Test?, June 2003. URL <http://www.lorenzgygax.ch/documents/introEtho.pdf>.
- [10] David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall: London, UK and CRC Press, Inc.: Boca Raton, FL, USA, 2nd/3rd edition, 2004. ISBN 0203489535, 0849331196, 1-58488-133-X, and 1-58488-440-1.
- [11] Lisa Lavoie Harlow, Stanley A. Mulaik, and James H. Steiger, editors. *What If There Were No Significance Tests?* Multivariate Applications Book Series. Lawrence Erlbaum Associates, Inc. (LEA): Mahwah, NJ, USA, August 1997. ISBN 0805826343.

- [12] Joel R. Levin. What If There Were No More Bickering about Statistical Significance Tests? *Research in the Schools (RITS)*, 5(2):43–53, 1998. URL <http://www.personal.psu.edu/users/d/m/dmr/sigtest/6mspdf.pdf>.
- [13] Gerard E. Dallal. The Little Handbook of Statistical Practice, July 16, 2008. URL <http://www.statisticalpractice.com/>.

```
@misc{W2011IOSTRFEE,
  author      = {Thomas Weise},
  title       = {Illustration of Statistical Test Results for Experiment Evaluation},
  publisher    = {it-weise.de (self-published): {Germany}},
  year        = {2011},
  month       = mar # {~2, },
  url         = {http://www.it-weise.de/documents/files/W2011IOSTRFEE.pdf},
  abstract    = {In this reference document, I provide some notes on how I
    (personally) think that the results of statistical tests involving
    \ensuremath{N>2} processes could be represented in a compact and
    easy-to-read way. The proposed method is a directed graph acyclic
    graph (DAG), which is based on the fact that the outcomes of tests
    always are at least a strict partial order which can be visualized
    as DAG.},
  keywords    = {Statistical Tests, Visualization, Directed Acyclic Graph, DAG,
    Mann-Whitney U Test, Java},
},
```