Comparison of Non-Deterministic Methods

STAMP Software

Éric Taillard

EIVD

University of Applied Sciences of Western Switzerland,

Yverdon-les-bains, Switzerland

BINARY COMPARISON (FOUND - NOT FOUND)

FANT Local search Iterations	FANT Optimum found	FANT CPU time	MMAS Local search Iterations	MMAS Optimum found	MMAS CPU time
25000	No	83.0	1335	Yes	4.4
778	Yes	2.7	26104	No	85.0
25000	No	83.0	26104	No	85.0
24658	Yes	81.9	26104	No	85.0
1052	Yes	3.5	173	Yes	0.6
19787	Yes	65.57	1389	Yes	4.5
784	Yes	2.61	26104	No	85.0
1009	Yes	3.35	26104	No	85.0
253	Yes	0.85	26104	No	85.0
5010	Yes	16.62	26104	No	85.0

Conditions of the experiments : Few runs (central limit theorem does not apply), Limited number of nodes (iterations) and/or limited CPU time.

Questions :

Is FANT significantly better than MMas (i. e. is a 8/10 rate of success significantly better than a 3/10 rate) ?

Is FANT significantly better after 10 seconds ? and, in general, after t seconds ($0.6 < t < 83$) ?

Statistical test for binary comparison

Classical U test :

Hypothesis : large samples (> 30)

Problem : Compare the relative frequencies f_a and f_b of Yes answers in two samples

Null hypothesis : $f_a = f_b$

Alternate hypothesis : $f_a > f_b$

Samples : (n_a runs, a Yes), (n_b, b)

Compute : $$U = \frac{a/n_a - b/n_b}{\sqrt{\frac{a/n_a \cdot (1 - a/n_a)}{n_a} + \frac{b/n_b \cdot (1 - b/n_b)}{n_b}}}$$

Reject the null hypothesis with confidence level α if $U > u_{1-\alpha}$

In other terms : reject the hypothesis that heuristic A is as good as B if the difference in the proportions of the Yes answer is too large.

NON STANDARD TEST FOR BINARY COMPARISON

Problem : Compare the relative frequencies f_a and f_b of Yes answers in two samples

Null hypothesis : $f_a = f_b = p$

Alternate hypothesis : $f_a > f_b$

Samples : (n_a runs, a Yes), (n_b, b)

Method :

Suppose that the Yes answer has the same probability $p = \dfrac{a+b}{n_a + n_b}$ to appear for both methods A and B

The probability P to observe a Yes answers or more for method A

b Yes answers or less for method B

is given by $$P = \sum_{i=a}^{n_a} \sum_{j=0}^{b} \binom{n_a}{i} \cdot p^i \cdot (1-p^i) \cdot \binom{n_b}{j} \cdot p^j \cdot (1-p^j)$$

Reject null hypothesis if $P < 1-\alpha$

For an iterative method, this test has to be repeated for any computational effort *t*

COMPARING THE QUALITY OF HEURISTIC METHODS

Which heuristic is best (minimize the objective function) ?

Heuristic S		Heuristic T		Answer	
Average : 100		Average : 101			S
Average : 100	Standard dev. : 6	Average : 100	Standard dev : 12	Today	S?
Average : 100	Standard dev. : 6	Average : 100	Standard dev : 12	Tomorrow	T?
Average : 100	Standard dev. : 6	Average : 101	Standard dev : 12	Tomorrow	T?

The right answer depends on the (unknown !) solution value distribution.

Examples : 10 runs of heuristics A, B, C

A : 98 98 98 98 98 98 98 98 98 118 Average : 100, Sd : 6

B : 96 96 96 96 96 96 96 96 96 136 Average : 100, Sd 12

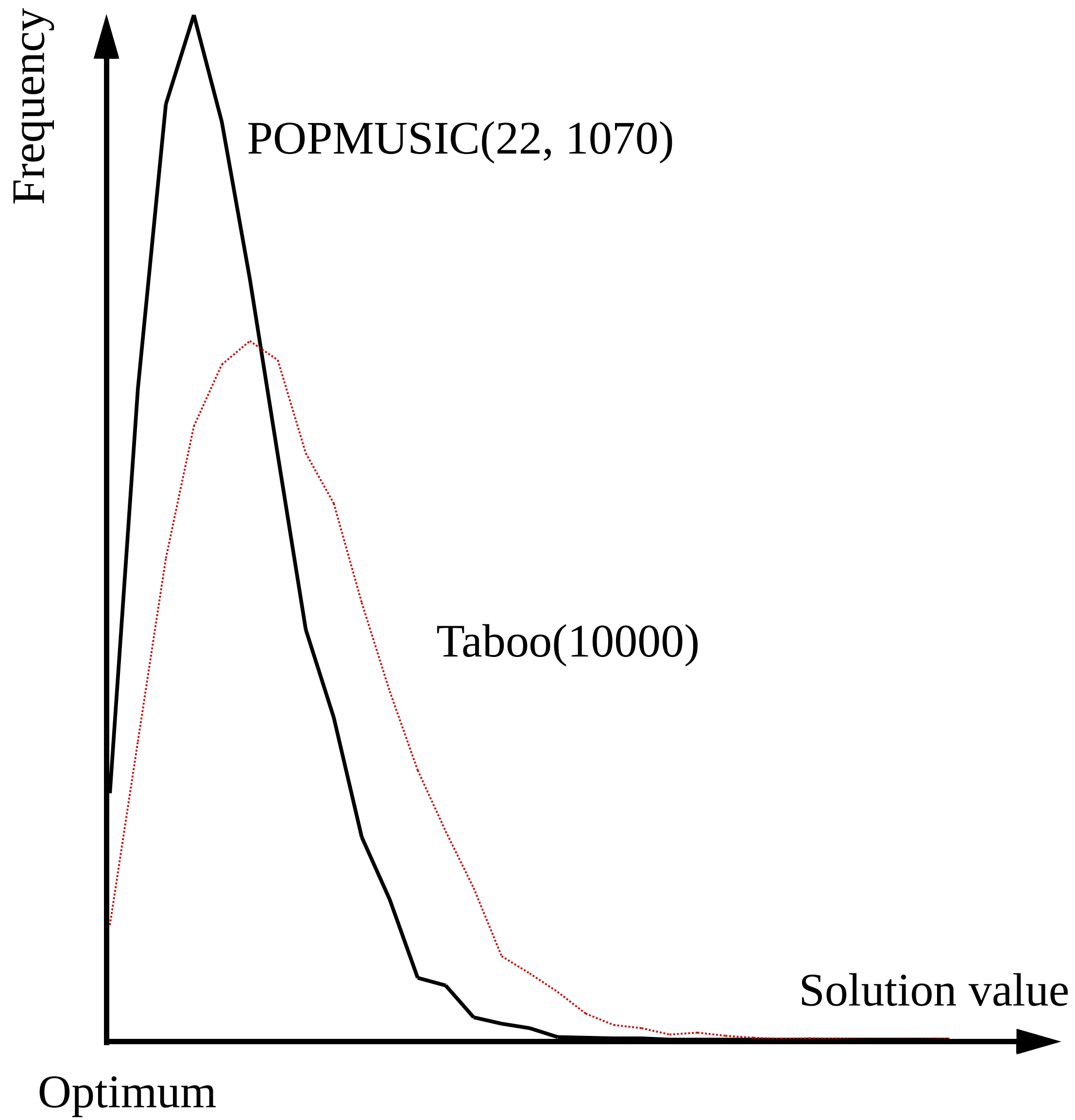
C : 97 97 97 97 97 97 97 97 97 137 Average : 101, Sd 12

P(one run of A better than one run of B) = P(A < B) : 0.1

2 independent runs : P(min(A, A) < min (C, C)) : 0.01

SOLUTION VALUE DISTRIBUTIONS

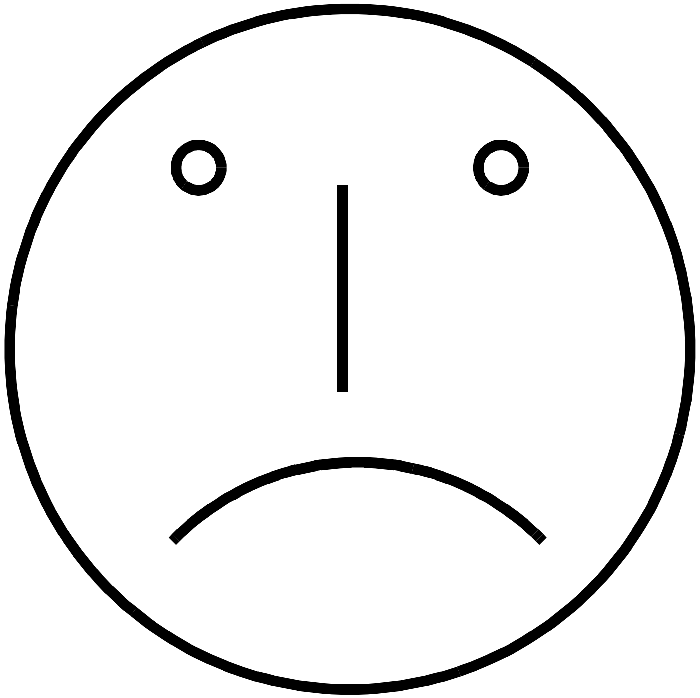
Not the same, not Gaussian ! Generally not known and cannot be determined.



STANDARD STATISTICAL TEST

Classical T test :

Hypothesis : Gaussian distributions (or large samples).



NON PARAMETRIC STATISTICAL TEST

Experiments : Run heuristics *A* and *B* n_a and n_b times for the same problem instance (long runs)

Question : Is heuristic *B* significantly worse than heuristic *A* for a given computational effort ?

Method :

Null hypothesis : Suppose that both methods *A* and *B* provide solutions with the same distribution.

Rank the $n_a + n_b$ runs of heuristics *A* and *B* by decreasing quality (for a given effort *t*)

Compute $T = \sum$ ranks heuristic *A*

The probability *P* to observe a $\sum$ ranks heuristic $A < T$ is given by

$P = f(T, n_a, n_b)$, a function that takes few pages of ada code

Reject null hypothesis at confidence level α if $P < 1-\alpha$

For a heuristic iterative method, this test has to be repeated for any computational effort *t*

Note : variants of this test are known under the name of Mann-Whitney test

CONCLUSIONS

Tell the whole story

Various heuristic methods.

Various problem instances.

Large span of computational effort (logarithmic scale).

Report negative experiments

Understanding the working principles of methods.

Graphical analysis of results

Distribute code or complete numerical results.

Ensure significance of results

Drawing conclusions supported by the results.

Help for parameter tuning