A Statistical Test for Comparing Success Rates

Éric D. Taillard*

*EIVD, University of Applied Sciences of Western Switzerland
Route de Cheseaux 1, Case postale
CH-1401 Yverdon-les-Bains, Switzerland

Eric.Taillard@eivd.ch

1 Introduction

Who was not once perplex when reading, in an article comparing two optimization methods, numerical results presented under the following form : "We have tested our optimization method A on a set of n problem instances from the literature and we succeeded in solving a of these instances. The concurrent method B succeeded in solving only b of these instances. However, it has to be noted that method B was tested only on m over the n instances". Indeed, the reader has no answer to the basic question : "Is a success rate of a/n significantly superior to a success rate of b/m ? ". Very often, the answer to this central question cannot be found in classical statistical tests, since the lasts require large sample sizes (at least about 15).

Nevertheless, in combinatorial optimization, problem instances sets are frequently smaller than 15. Intuitively, the reader is perfectly convinced that a method A that succeeded in solving all 10 problem instances of a given set is better than a method B that solved none of them. On the contrary, the reader will not be really convinced if the problem set contains only 3 instances. However, supposing that the problem instances have been chosen independently from the solving methods, it can be shown that a 3/3 rate of success is significantly larger (with a confidence level higher than 98%) than a 0/3 rate of success.

It could be argued that larger problem sets must be used, so that standard statistical sets could be applied. Unfortunately this is not always possible. First, for real problems, collecting data for a single instance may take several weeks. Second, there are classical problem instances libraries (ORLIB [2], QAPLIB [3], TSPLIB [4]) that seldom propose more than 10 instances for a given problem size. Third, it can be noted that many optimization methods are very time consuming (for instance the code for the quadratic assignment problem of [1] which took the equivalent of 7 years CPU time on a sequential computer for solving instance Nug30). In such a case, it might be more interesting to estimate the performances of a method on few large problem instances than on a multitude of toys-instances.

2 Standard statistical test

The comparison of rate of successes for two populations A and B is traditionally done as follows : Let p_a (respectively p_b) be the probability of success of population A (respectively population B) and random samples of size n (respectively m) are taken for the experiment. Then, the statistic $U = X_a/n - X_b/m$ (where X_a and X_b are random variables associated with successful experiments in populations A and B) has the mean $p_a - p_b$ and variance $p_a \cdot q_a/n + p_b \cdot q_b/m$, where $q_a = 1 - p_a$ and $q_b = 1 - p_b$. For conducting a statistical test, the null hypothesis is $p_a = p_b = p$, so that U should have a mean of 0.

If both n and m are large, U is approximately normally distributed and the common probability p can be estimated by $\hat{p} = \frac{a+b}{m+n}$, where a and b are the number of successes observed in populations A and B. The null hypothesis will be rejected (and p_a will be considered to be higher than p_b) if $\Phi(\hat{u}) > \alpha$, where α is the confidence level, Φ is the cumulative standard normal distribution and

$$\hat{u} = \frac{a/n - b/m}{\sqrt{\frac{a+b}{n+m} \cdot \frac{n+m-a-b}{n+m} \cdot (1/n + 1/m)}}$$

Practically, if both n and m are higher than 14, the null hypothesis can be reasonably rejected if $\hat{u} > 1.65(\alpha = 95\%)$, $\hat{u} > 2.06(\alpha = 98\%)$, $\hat{u} > 2.33(\alpha = 99\%)$, $\hat{u} > 2.58(\alpha = 99.5\%)$, $\hat{u} > 3.09(\alpha = 99.9\%)$.

3 A non parametric statistical test

In order to answer the question : "Does a observation of a criterion over a sample of size n represent a rate higher than b occurrences over a sample of size m ? ", it can be proceeded as follows :

Null hypothesis : Let us suppose that the (unknown) rate p of occurrence of the criterion is the same for both sample (i.e. $p_a = p_b = p$). Under the null hypothesis, the probability $S(p, a, n, b, m)$ to observe a successes or more in the first population (of size n) and b successes or less in the second population (of size m) is given by the product of two binomial distributions ($C_i^n = \frac{n!}{i! \cdot (n-i)!}$ and $C_j^m = \frac{m!}{j! \cdot (m-j)!}$ are the binomial coefficients) :

$$S(p, a, n, b, m) = (\sum_{i=a}^{n} C_i^n \cdot p^i \cdot (1-p)^{n-i}) \cdot (\sum_{j=0}^{b} C_j^m \cdot p^j \cdot (1-p)^{m-j})$$

Alternate hypothesis : $p_a > p_b$ i.e. the success rate of method A is higher than the success rate of method B.

The null hypothesis has to be rejected with a confidence level α (and the alternate hypothesis accepted, i.e. an a/n rate will be considered higher than a b/m rate) if

$$\max_{0<p<1} S(p, a, n, b, m) \leq 1 - \alpha$$

4 Examples

Let us suppose that all n observations from the first sample where successes and all m observations from the second sample where failures (i.e. $a = n$ and $b = 0$). Supposing that both populations have the same probability of success, $S(p, n, n, 0, m) = p^n \cdot (1-p)^0 \cdot p^0 \cdot (1-p)^m = p^n \cdot (1-p)^m$.

The probability $\hat{p}$ that maximizes $S(p, n, n, 0, m)$ is given by solving the equation :

$$\frac{dS(p, n, n, 0, m)}{dp} = np^{n-1} \cdot (1-p)^m - mp^n \cdot (1-p)^{m-1} = 0$$

For the special case $a = n$ and $b = 0$, the pooled estimate $\hat{p} = \frac{a+b}{m+n}$ is therefore the value that maximizes $S(p, a, n, b, m)$ over p. For instance, if $n = 3$ and $m = 2$, $S(3/5, 3, 3, 0, 2) = 108/3125 < 5\%$. So a success rate of 3/3 is significantly higher (with confidence level of 95%) than a success rate of 0/2.

Unfortunately, for arbitrary values of a, n, b and m, the pooled estimate is *not* the value that maximizes $S(p, a, n, b, m)$ over p. For instance, $S(3/7, 3, 4, 0, 3) < 4/100$ and $S(\frac{6-2\sqrt{2}}{7}, 3, 4, 0, 3) > 4/100$. This means that testing if a rate of 3/4 is significantly higher than a rate of 0/3 with a confidence level of 96% whould lead to an erroneous conclusion if the pool estimate is used. In general, the analytic expression of $\hat{p}$ is at least hard to be found in practice. Therefore, we have numerically estimated $\hat{p}$ and provide in Table 1 (and, respectively in Table 2), for various values of n and m and for a confidence level of 95% (respectively 99%), the most extreme couples (a, b) for which an a/n rate of success is higher than a b/m rate.

Reading the tables Let us suppose that the observed success rate of an optimization method A is 6/10 and the observed success rate of method B is 1/9 (meaning that $a = 6$, $n = 10$, $b = 1$, $m = 9$). In Table 1, entry $n = 10$ and $m = 9$ contains the couple (5,1), meaning that a 5/10 success rate is significantly higher than a 1/9 success rate at 95% confidence level. Since the success rate $6/10 > 5/10$ it can be deduced that method A is significantly better than method B (at 95% confidence level).

5 Conclusions

This article presents a non-parametric statistical test that is very interesting for those who want to compare different heuristic algorithms that do not necessarily end with feasible (or satisfying) solutions. This test has been specially designed for working with very small sample sizes, meaning that a substantial computational effort can be saved when conducting numerical experiments.

When the sample sizes are lower than 15, standard statistical tests for comparing the success rates of two populations cannot be validly used. We have indeed observed that the standard statistical test — abusively applied — provides results that are erroneous. So it is for very high confidence rates, even if sample sizes are larger than 15. Therefore, a non parametric

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2		(3,0)	(4,0)	(5,0)	(5,0)	(6,0)	(7,0)	(7,0)	(8,0)	(9,0)	(9,0)	(10,0)	(11,0)	(11,0)
													(14,1)	(15,1)
3	(2,0)	(3,0)	(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(8,0)	(9,0)	(9,0)
				(5,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(10,1)	(11,1)	(12,1)	(12,1)	(13,1)
4	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(8,0)
		(3,1)	(4,1)	(5,1)	(5,1)	(6,1)	(7,1)	(7,1)	(8,1)	(9,1)	(9,1)	(10,1)	(11,1)	(11,1)
					(6,2)	(7,2)	(8,2)	(9,2)	(10,2)	(11,2)	(12,2)	(12,2)	(13,2)	(14,2)
5	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)
		(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(9,1)	(10,1)
			(4,2)	(5,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(10,2)	(10,2)	(11,2)	(12,2)	(12,2)
							(8,3)	(9,3)	(10,3)	(11,3)	(12,3)	(13,3)	(14,3)	(15,3)
6	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)	(6,0)
	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)	(8,1)	(9,1)
		(3,2)	(4,2)	(5,2)	(5,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(9,2)	(10,2)	(11,2)	(11,2)
				(5,3)	(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(12,3)	(12,3)	(13,3)
									(10,4)	(11,4)	(12,4)	(13,4)	(14,4)	(15,4)
7	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(7,1)	(8,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)	(10,2)	(10,2)
			(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(11,3)	(12,3)
					(6,4)	(7,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)	(13,4)	(14,4)
										(11,5)	(12,5)	(13,5)	(14,5)	(15,5)
8	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)	(7,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)	(9,2)
		(3,3)	(4,3)	(5,3)	(5,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(10,3)	(11,3)
				(5,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)	(13,4)
						(7,5)	(8,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)	(13,5)	(14,5)
												(13,6)	(14,6)	(15,6)
9	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)
	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(10,3)
			(4,4)	(5,4)	(6,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)
				(5,5)	(6,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)	(13,5)
							(8,6)	(9,6)	(10,6)	(11,6)	(12,6)	(13,6)	(14,6)	(14,6)
														(15,7)
10	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)
	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)
		(3,4)	(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)	(11,4)
			(4,5)	(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)
					(6,6)	(7,6)	(8,6)	(9,6)	(10,6)	(10,6)	(11,6)	(12,6)	(13,6)	(14,6)
								(9,7)	(10,7)	(11,7)	(12,7)	(13,7)	(14,7)	(15,7)
11	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)
	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(7,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(7,3)	(8,3)	(8,3)	(9,3)
		(3,4)	(4,4)	(5,4)	(5,4)	(6,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)
			(4,5)	(5,5)	(6,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)	(12,5)
				(5,6)	(6,6)	(7,6)	(8,6)	(8,6)	(9,6)	(10,6)	(11,6)	(11,6)	(12,6)	(13,6)
						(7,7)	(8,7)	(9,7)	(10,7)	(11,7)	(11,7)	(12,7)	(13,7)	(14,7)
									(10,8)	(11,8)	(12,8)	(13,8)	(14,8)	(15,8)
12	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(5,1)
	(2,2)	(3,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(6,2)	(7,2)	(7,2)
	(2,3)	(3,3)	(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(8,3)	(8,3)
		(3,4)	(4,4)	(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(8,4)	(9,4)	(9,4)	(10,4)
		(3,5)	(4,5)	(5,5)	(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)
			(4,6)	(5,6)	(6,6)	(7,6)	(7,6)	(8,6)	(9,6)	(9,6)	(10,6)	(11,6)	(11,6)	(12,6)
				(5,7)	(6,7)	(7,7)	(8,7)	(9,7)	(9,7)	(10,7)	(11,7)	(12,7)	(12,7)	(13,7)
						(7,8)	(8,8)	(9,8)	(10,8)	(11,8)	(12,8)	(12,8)	(13,8)	(14,8)
										(11,9)	(12,9)	(13,9)	(14,9)	(15,9)
13	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)
	(2,1)	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)
	(2,2)	(3,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(6,2)	(7,2)
	(2,3)	(3,3)	(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(8,3)
		(3,4)	(4,4)	(4,4)	(5,4)	(5,4)	(6,4)	(6,4)	(7,4)	(7,4)	(8,4)	(8,4)	(9,4)	(9,4)
		(3,5)	(4,5)	(5,5)	(5,5)	(6,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(9,5)	(10,5)	(10,5)
			(4,6)	(5,6)	(6,6)	(6,6)	(7,6)	(8,6)	(8,6)	(9,6)	(10,6)	(10,6)	(11,6)	(11,6)
				(5,7)	(6,7)	(7,7)	(7,7)	(8,7)	(9,7)	(10,7)	(10,7)	(11,7)	(12,7)	(13,7)
					(6,8)	(7,8)	(8,8)	(9,8)	(10,8)	(10,8)	(11,8)	(12,8)	(13,8)	(13,8)
							(8,9)	(9,9)	(10,9)	(11,9)	(12,9)	(13,9)	(14,9)	(14,9)
											(12,10)	(13,10)	(14,10)	(15,10)

Table 1: Couples (a, b) for which a success rate $\geq a/n$ is significantly higher than a successes rate b/m, for a confidence level of 95%.

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2						(7,0)	(8,0)	(9,0)	(10,0)	(11,0)	(12,0)	(12,0)	(13,0)	(14,0)
3			(4,0)	(5,0)	(6,0)	(7,0)	(7,0)	(8,0)	(9,0)	(9,0)	(10,0)	(11,0)	(11,0)	(12,0)
											(12,1)	(13,1)	(14,1)	(15,1)
4		(3,0)	(4,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(8,0)	(8,0)	(9,0)	(9,0)	(10,0)	(11,0)
					(6,1)	(7,1)	(8,1)	(9,1)	(10,1)	(11,1)	(11,1)	(12,1)	(13,1)	(14,1)
5		(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(8,0)	(8,0)	(9,0)	(9,0)
				(5,1)	(6,1)	(7,1)	(7,1)	(8,1)	(9,1)	(10,1)	(10,1)	(11,1)	(12,1)	(12,1)
								(9,2)	(10,2)	(11,2)	(12,2)	(13,2)	(14,2)	(14,2)
6		(3,0)	(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(8,0)	(8,0)	(9,0)
			(4,1)	(5,1)	(6,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(9,1)	(10,1)	(11,1)	(11,1)
					(6,2)	(7,2)	(8,2)	(9,2)	(10,2)	(10,2)	(11,2)	(12,2)	(13,2)	(13,2)
										(11,3)	(12,3)	(13,3)	(14,3)	(15,3)
7	(2,0)	(3,0)	(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(7,0)	(8,0)
			(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(9,1)	(10,1)	(10,1)
				(5,2)	(6,2)	(7,2)	(8,2)	(8,2)	(9,2)	(10,2)	(10,2)	(11,2)	(12,2)	(12,2)
							(8,3)	(9,3)	(10,3)	(11,3)	(12,3)	(12,3)	(13,3)	(14,3)
												(13,4)	(14,4)	(15,4)
8	(2,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(6,0)	(7,0)	(7,0)
		(3,1)	(4,1)	(4,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)	(8,1)	(9,1)	(9,1)	(10,1)
			(4,2)	(5,2)	(6,2)	(6,2)	(7,2)	(8,2)	(8,2)	(9,2)	(10,2)	(10,2)	(11,2)	(12,2)
					(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(12,3)	(12,3)	(13,3)
								(9,4)	(10,4)	(11,4)	(12,4)	(13,4)	(14,4)	(14,4)
														(15,5)
9	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(6,0)	(7,0)
		(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)	(8,1)	(9,1)	(9,1)
			(4,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(9,2)	(10,2)	(10,2)	(11,2)
				(5,3)	(6,3)	(7,3)	(8,3)	(8,3)	(9,3)	(10,3)	(10,3)	(11,3)	(12,3)	(12,3)
						(7,4)	(8,4)	(9,4)	(10,4)	(11,4)	(11,4)	(12,4)	(13,4)	(14,4)
									(10,5)	(11,5)	(12,5)	(13,5)	(14,5)	(15,5)
10	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(6,0)
		(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)	(8,1)	(9,1)
			(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(8,2)	(8,2)	(9,2)	(9,2)	(10,2)	(10,2)
				(5,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(10,3)	(11,3)	(12,3)
					(6,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(11,4)	(12,4)	(12,4)	(13,4)
							(8,5)	(9,5)	(10,5)	(11,5)	(12,5)	(13,5)	(13,5)	(14,5)
											(12,6)	(13,6)	(14,6)	(15,6)
11	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)
		(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)	(8,1)
		(3,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)	(9,2)	(10,2)
			(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(8,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(11,3)
				(5,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)	(12,4)
						(7,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)	(12,5)	(13,5)	(13,5)
								(9,6)	(10,6)	(11,6)	(12,6)	(13,6)	(14,6)	(14,6)
												(13,7)	(14,7)	(15,7)
12	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)	(6,0)
		(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(7,1)	(8,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)	(9,2)
			(4,3)	(5,3)	(5,3)	(6,3)	(7,3)	(7,3)	(8,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)
				(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)
					(6,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(11,5)	(12,5)	(13,5)
							(8,6)	(9,6)	(10,6)	(11,6)	(12,6)	(12,6)	(13,6)	(14,6)
									(10,7)	(11,7)	(12,7)	(13,7)	(14,7)	(15,7)
													(14,8)	(15,8)
13	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)	(7,1)	(7,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)
			(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(10,3)
			(4,4)	(5,4)	(6,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(9,4)	(10,4)	(11,4)	(11,4)
				(5,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)	(12,5)	(12,5)
						(7,6)	(8,6)	(9,6)	(10,6)	(10,6)	(11,6)	(12,6)	(13,6)	(13,6)
							(8,7)	(9,7)	(10,7)	(11,7)	(12,7)	(13,7)	(13,7)	(14,7)
										(11,8)	(12,8)	(13,8)	(14,8)	(15,8)
14	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)	(7,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(8,2)
		(3,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(8,3)	(9,3)	(9,3)	(10,3)
			(4,4)	(5,4)	(6,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)	(11,4)
				(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(11,5)	(12,5)
					(6,6)	(7,6)	(8,6)	(9,6)	(9,6)	(10,6)	(11,6)	(11,6)	(12,6)	(13,6)
						(7,7)	(8,7)	(9,7)	(10,7)	(11,7)	(11,7)	(12,7)	(13,7)	(14,7)
								(9,8)	(10,8)	(11,8)	(12,8)	(13,8)	(14,8)	(14,8)
											(12,9)	(13,9)	(14,9)	(15,9)
15	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)
		(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(8,3)	(9,3)	(9,3)
			(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)
				(5,5)	(6,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)	(11,5)
				(5,6)	(6,6)	(7,6)	(8,6)	(8,6)	(9,6)	(10,6)	(10,6)	(11,6)	(12,6)	(12,6)
						(7,7)	(8,7)	(9,7)	(10,7)	(11,7)	(11,7)	(12,7)	(13,7)	(13,7)
							(8,8)	(9,8)	(10,8)	(11,8)	(12,8)	(13,8)	(14,8)	(14,8)
									(10,9)	(11,9)	(12,9)	(13,9)	(14,9)	(15,9)
												(13,10)	(14,10)	(15,10)

Table 2: Couples (a, b) for which a success rate $\geq a/n$ is significantly higher than a successes rate b/m, for a confidence level of 99%.

test has been developed. This test is more accurate and can be applied for any sample sizes, but it requires relatively heavy computations. So, pre-computed values for 95% and 99% confidence levels have been tabulated in the present article. The computation of confidence levels can also be done *online* at the URL : **http://ina.eivd.ch/projects/stamp/**

When the sample sizes are at least 15, we have observed that the probability of rejecting the null hypothesis when the last is true for the standard test almost always over estimates the corresponding value obtained with the non parametric test. This means that the standard test very seldom reject the null hypothesis when it has to be accepted, according to the non parametric test. Let us mention that, very often, the standard test strongly over estimates the value of the probability of the null hypothesis, meaning that the non parametric test proposed is more powerful than the standard one.

6 Acknowledgements

The author would like to thank F. Taillard and J. Zuber for comments and discussions on early versions of the article as well as A. Løkketangen for asking the embarrassing questions about the comparison of methods for the satisfiability problem that have led to the development of the present statistical test. The online implementation for computing the statistical test is due to Ph. Wälti. The present work is supported by the strategic founds of the Applied University of Western Switzerland, grant LQF01-03.

References

[1] K.M. Anstreicher, N.W. Brixius, J.-P. Goux and J. Linderoth, "Solving large quadratic assignment problems on computational grids", to appear in *Mathematical Programming, Series B*, 2001. Currently available on the Web from http://www.biz.uiowa.edu/faculty/anstreicher/mwqap.ps

[2] J.E. Beasley, "OR-Library: distributing test problems by electronic mail", *Journal of the Operational Research Society 41(11)*, 1990, pp. 1069–1072. http://mscmga.ms.ic.ac.uk/info.html

[3] R.E. Burkard, S.E. Karisch and F. Rendl, "QAPLIB A Quadratic Assignment Problem Library", *Journal of Global Optimization 10*, pp. 391–403, 1997. http://www.opt.math.tu-graz.ac.at/qaplib

[4] G. Reinelt, "TSPLIB : sample instances for the TSP (and related problems) from various sources and of various types". Available on the web : http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/index.html