Few statistical tests for proportions comparison

Éric D. Taillard, Philippe Waelti, Jacques Zuber

University of Applied Sciences of Western Switzerland
EIVD campus at Yverdon
Route de Cheseaux 1, case postale
CH-1401 Yverdon-les-Bains, Switzerland

Abstract

This article review few statistical tests for comparing proportions. These statistical tests are presented in a comprehensive way, so that OR practitionners can easily understand them and use them correctly. A non-parametric test is developped and shown to be more powerful than other classical tests of the literature.

1 Introduction

In operations research, comparing two solution methods each other is frequently needed. This is in particular the case when one wants to tune the parameters of an algorithm. In this case, one wants to know wether a given parameter setting is better than another one. In practice, for identifying the best setting, there are several approaches. Without being exhaustive common techniques are the following :

1. In the context of optimization, a set of problem instances is solved with both methods that have to be compared. Then, the average, standard deviation (an eventually other measures such as median, minimum, maximum, skewness, kurtosis, etc.) of the solution values obtained are computed.
2. In the context of solving problems exactly, the average, standard deviation, etc. of the computational effort needed to obtain the optimum solution are computed.
3. The maximual computational effort is fixed, as well as a goal reach. The number of times each method reach the goal is counted.

Naturally, there are many variants and other statistics that can be collected. In the first comparison technique, the computational effort is not taken into account. Either the last is very small, or both methods requires approximately the same computational effort.

Very often in practice, the measures that are computed in the first and second comparison techniques quoted above are very primitive. Sometimes they are limited to the only average. This is evidently very insufficient for stating that a solution method is statistically better than another one.

When the standard deviation is computed in addition to the average, it is possible to perform a valid statistical test, as soon as the hypothesis of *normal distribution* of the data is satisfied. The last hypothesis is far from being always satisfied. For instance, an optimization technique that frequently finds globally optimal solutions has a distribution with a truncated queue, since it is impossible to go beyond the optimum. This situation is illustrated on Figure 1 that provides the empirical distributions of solutions values obtained for two nondeterministic optimization techniques (Robust taboo search[Taillard(1991)] and POPMUSIC[Taillard & Voss(2002)]) for a problem instance of turbine runner balancing.

This figure shows clearly that the distributions are asymetrical, left truncated (this is a minimization problem; the vertical axis is placed on a lower bound to the optimum) and that both distribution functions are different. Therefore, the estimation of a parameter (the average) of an a-priori unknown distribution function is not evident. Moreover, the probability that the estimation of the average for standing between two bounds should be given, which seems to be a difficult task to be undertaken.

For this reason, nonparametric statistics have been developped. Indeed, they are based on weaker hypothesis. When the third comparison approach quoted above is used (counting the number of successes), the sign test[Arbuthnott(1710)] (see, e.g.[Conover(1999)]) could be convenient. The present article develops a nonparametric statistical test that is more powerful than the sign test for comparing proportions. This test is perhaps a new one, since it is not developped in the literature consulted, although it cannot be excluded that it appears somewhere, since there is a huge amount of articles dealing with contingency tables.

2 Comparing proportions

The central problem treated by the present article is the following: Let us suppose that two populations A and B are governed by binomial distributions, i.e. the probability of success of an occurence of A (respectively: B) is given by p_a (respectively: p_b). From the OR user point of view, it is considered that the result of the execution of a method is a random variable. Indeed, either the method is nondeterministic which is typically the case of simulated

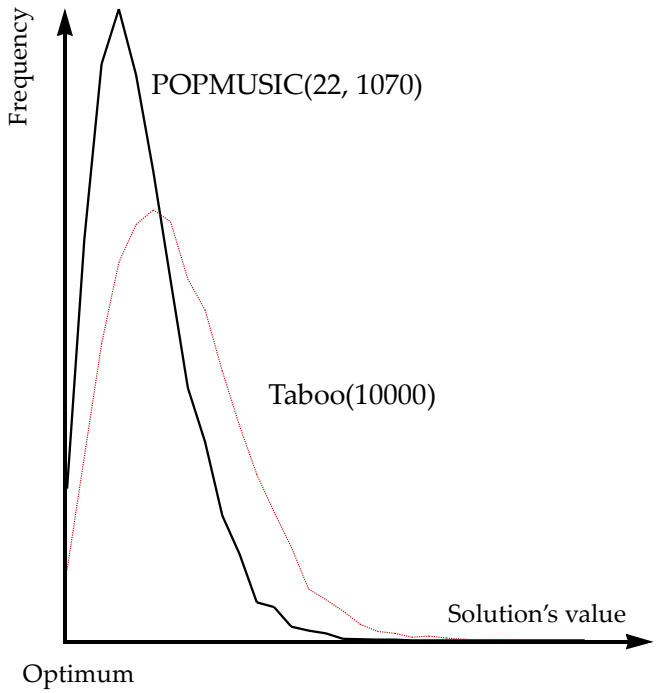

Figure 1: Empirical distributions of solution values obtained by two non-deterministic methods (POPMUSIC and taboo), obtained by solving a large number of times the same problem instance.

annealing, or the problem data can be viewed as random, the user being not able to influence them. So, it is supposed that Method A (respectively: Method B) has a probability of p_a (respectively: p_b) to be successfull.

The user would like to use the method which success probability is the higher. Ideally, the user wants to know p_a and p_b to make a choice. Unfortunately, these probabilities are unknown. The user could try to estimate them empirically. In the following, it is considered that the user has proceeded to the following experiment:

Sampling Method A (respectively Method B) has been run n_a (respectively: n_b) times and was successfull a (respectively: b) times.

2.1 Classical parametric approaches

The classical approach (based on the central limit theorem) for comparing two proportions is the following: Let X_a (respectively X_b) be the random variable associated to the number of successes of Method A (respectively: Method B). Then, the average of the random variable $D = X_a/n_a - X_b/n_b$ is $d = p_a - p_b$ and the variance of D is $\sigma_D^2 = p_a \cdot q_a/n_a + p_b \cdot q_b/n_b$ where $q_a = 1 - p_a$ and $q_b = 1 - p_b$. If n_a and n_b are large enough (an empirical rule often used is $min(n_a \cdot p_a \cdot q_a, n_b \cdot p_b \cdot q_b) > 5$), then D is approximately normally distributed.

In order to compare the success rates of methods A and B, one makes the following

Null hypothesis (one-sided test) Probability p_a is lower or equal to probability p_b, i.e. $d = p_a - p_b \leq 0$.

Alternative hypothesis $p_a > p_b$.

We recall here that the principle of an hypothesis statistical test is to estimate the probability of the observations done in case the null hypothesis is true. If this probability is higher than a significance level α then it cannot be excluded that the null hypothesis is true. α is set to a small value, typically 0.05 or 0.01. If the probability of the null hypothesis is lower than α, then the Alternative hypothesis (the logical negation of the null hypothesis) is accepted. In conducting an hypothesis statistical test, the null hypothesis is chosen in such a way that it is felt not to be true. So, in the above mentionned hypothesis, it can be assumed that the experiment has shown $a/n_a > b/n_b$. Note that there is another symmetrical one-sided test with null hypothesis $p_b - p_a \leq 0$. This other test is equivalent to those above with roles of samples A and B inverted. Since probabilities p_a and p_b are unknown, it is searched for probabilities that maximize the probability of null hypothesis to be true. This maximum occurs for

$\hat{p}_a = \hat{p}_b = \hat{p}$. It is considered that one can take the pooled estimate: $\hat{p} = \frac{a+b}{n_a+n_b}$. The value observed for d is given by $\hat{d} = a/n_a - b/n_b$ and the variance can be estimated by $\hat{s}^2 = \hat{p} \cdot \hat{q}/n_a + \hat{p} \cdot \hat{q}/n_b$, where $\hat{q} = 1 - \hat{p}$.

The distribution of the null hypothesis for large n_a and n_b is $N(0, \sigma_D^2)$. So, the null hypothesis is not plausible, at significance level α if $\Phi(\hat{d}/\hat{s}) < \alpha$, where Φ is the cumulative normal distribution. In practice, the null hypothesis is rejected at significance levels :

- $\alpha = 5\%$ if $\hat{d}/\hat{s} > 1.645$
- $\alpha = 1\%$ if $\hat{d}/\hat{s} > 2.326$
- $\alpha = 0.1\%$ if $\hat{d}/\hat{s} > 3.09$

The above mentionned statistical test is a simplification of the "Chi-square Test for Difference in Probabilities, 2×2 contigency table" [Conover(1999)]. Indeed, in case of the two-sided test :

Null hypothesis $p_a = p_b$

Alternative hypothesis $p_a \neq p_b$

Then it can be shown that, for large n_a and n_b, the distribution of the test statistic :

$$T = \frac{(n_a + n_b) \cdot (an_a - bn_b)^2}{n_a \cdot n_b \cdot (a+b) \cdot (n_a + n_b - a - b)}$$

i.e. $(\frac{\hat{d}}{\hat{s}})^2$, can be approximated, under the null hypothesis, by the Chi-square distribution with 1 degree of freedom.

In practice, the null hypothesis is rejected (and the alternative hypothesis $p_a \neq p_b$ is accepted) at significance levels :

- $\alpha = 5\%$ if $T > 3.841$
- $\alpha = 1\%$ if $T > 6.635$
- $\alpha = 0.1\%$ if $T > 10.83$

The interested reader may find more information about these approaches in [Cramér(1946), Harkness & Katz(1964), Ott & Free(1969)].

2.2 McNemar test for signifiance of changes

In many situations, both samples are of same size since one tries to test the effect of a treatment by making an experience before and an experience after the treatment. So, one has pairwise data that represents the condition of the subject before and after the treatment. This situation occurs in the operations research when one wants to know wether Method A is signifiantly more successful than Method B by running both methods on the same data set.

Let a' be the number of times pair (success, failure) has been observed (i.e. success of Method A and failure for Method B) and b' be the number of times pair (failure, success) has been observed over the $n'_a = n'_b = n$ observations. Thus experiments that provide the same result for both methods (success, success) or (failure, failure) are eliminated.

Null hypothesis

- Two-sided test : $P(failure, success) = P(success, failure) = 1/2$
- One-sided test : $P(failure, success) \leq P(success, failure)$

Alternative hypothesis

- Two-sided test : $P(failure, success) \neq P(success, failure)$
- One-sided test : $P(failure, success) > P(success, failure)$

Decision rule The null hypothesis is rejected at significance level α if :

- Two-sided test : $\frac{1}{2^n} \cdot \sum_{i=0}^{a'} C_i^n < \alpha/2$ or if $\frac{1}{2^n} \cdot \sum_{i=0}^{a'} C_i^n > 1 - \alpha/2$, where $C_i^n = \frac{n!}{i! \cdot (n-i)!}$
- One-sided test : $\frac{1}{2^n} \cdot \sum_{i=0}^{a'} C_i^n < \alpha$

The advantage of McNemar test is that it can be applied to any sample size, since it is based on the binomial distribution and not on the central limit theorem.

3 A new test for comparing proportions

The drawback of McNemar test is that pairwise data are required. In practice it is not always possible to have pairwise data. For instance, let us suppose that Method B was run on n_b problem instances randomly generated. The rules for

problem generation are perfectly known, but the n_b instances themselves have not been published. So, the designer of Method A, who wants to compare his method to Method B, can run his method as many times as he wants (n_a times). However, if the code of Method B is not available, he only knows that Method B was successfull b times over n_b runs. If n_b is not large, then the standard test cannot be validly applied. Moreover, if the designer of Method A chooses $n_b = n_a$, then McNemar test might be not significant, even if Method B was always sucessfull ($b = n_b$), whilst he could choose a larger value for n_a (and thus getting significant differences). Therefore, we developped a new statistical test for comparing proportions. In the remaining of the paper we make the following :

Assumptions

- The size of sample A is n_a ; a successes and $n_a - a$ failures have been observed.
- The size of sample B is n_b ; b successes and $n_b - b$ failures have been observed.
- Observations are mutually independent.
- The probability p_a (respectively: p_b) to have a success for population A (respectively: B) doesn't depends on the observations. Either $p_a < p_b$ or $p_b < p_a$ or $p_a = p_b$ for all observations (unbiased sample).

3.1 Two-sided test

The two-sided test is based on the following hypothesis :

Null hypothesis $p_a = p_b = p$

Alternative hypothesis $p_a \neq p_b$

Under the null hypothesis, the probability T to observe a successes in a sample of size n_a and b successes in a sample of size n_b is given by :

$$T = \frac{C_a^{n_a} \cdot p^a \cdot (1-p)^{n_a - a} \cdot C_b^{n_b} \cdot p^b \cdot (1-p)^{n_b - b}}{C_{a+b}^{n_a+n_b} \cdot p^{a+b} \cdot (1-p)^{n_a+n_b-a-b}} = \frac{C_a^{n_a} \cdot C_b^{n_b}}{C_{a+b}^{n_a+n_b}}$$

This probability does not depend on p, but only on n_a, n_b, a, b. Thus, T is very simple to compute.

Decision rule The null hypothesis ($p_a = p_b$) is rejected at significance level α if $T < \alpha$.

The test is in relation with Fishers's exact test ([Finney(1948), Robertson(1960)], see also [Gail & Gart(1973), Garside & Mack(1976), McDonald et al.(1977)])

3.2 One-sided test

For this test, let us suppose that the user wants to show that sample A has a higher success proportion than population B (e.g. the user has observed $a/n_a > b/n_b$). This can be assumed without loss of generality since the role of samples A and B can be reversed if the test is wanted to be conducted in the other way.

Null hypothesis $p_a \leq p_b$

Alternative hypothesis $p_a > p_b$

The probability S to observe a successes or more over n_a observations and b successes or less over n_b observations is given by :

$$S = (\sum_{i=a}^{n_a} C_i^{n_a} \cdot p_a^i \cdot (1-p_a)^{n_a-i}) \cdot (\sum_{j=0}^{b} C_j^{n_b} \cdot p_b^j \cdot (1-p_b)^{n_b-j})$$

This probability depends on proportions p_a and p_b which are unknown. Since the null hypothesis is wanted to be rejected with the highest security, probability S must be maximized over p_a and p_b, subject to the constrain that the null hypothesis is satisfied, i.e. $p_a \leq p_b$. It is clear that the maximum occurs for $p_a = p_b$ if $a/n_a > b/n_b$.

Decision rule The null hypothesis is rejected at significance level α if :

$$\hat{S} = max_{0<p<1}(\sum_{i=a}^{n_a} \sum_{j=0}^{b} C_i^{n_a} \cdot C_j^{n_b} \cdot p^{i+j} \cdot (1-p)^{n_a+n_b-i-j}) < \alpha$$

3.2.1 Examples

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

The probability $\hat{p}$ that maximizes S is given by solving the equation :

$$\frac{dS}{dp} = n_a \cdot p^{n_a-1} \cdot (1-p)^{n_b} - n_b \cdot p^{n_a} \cdot (1-p)^{n_b-1} = 0$$

.

For the special case $a = n_a$ and $b = 0$, the pooled estimate $\hat{p} = \frac{a+b}{n_a+n_b}$ is therefore the value that maximizes S over p. For instance, if $n_a = 3$ and $n_b = 2$, $S = 108/3125 < 5\%$. 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_a, b and n_b, the pooled estimate is *not* the value that maximizes S over p. For instance, for $a = 3$, $n_a = 4$ $b = 0$ and $n_b = 3$, $S < 4/100$ for $p = 3/7$ and $S > 4/100$ for $p = \frac{6-2\sqrt{2}}{7}$.

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 pooled estimate is used.

Although the difference in S values for the above example is not very large, the pooled estimate of p may underestimate by more than 1/3 the S value regarding to the maximum of S over p. This is exemplified by a success rate of 4/4 compared to a success rate of 56/100. The pooled estimated would provide a S value lower than 4.5% while there is a value of p that provides a value of S near 6%.

3.2.2 Computation of decision rule

In general, the analytic expressions of $\hat{p}$ and $\hat{S}$ are at least hard to be found in practice. Therefore, we have numerically estimated $\hat{S}$ and provide in Table 1 (and, respectively, in Table 2), for various values of n_a and n_b and for a significance level of 5% (respectively 1%), the most extreme couples (a, b) for which it is not plausible that an a/n_a rate of successes is lower than a b/n_b rate.

Reading the tables Due to the large number of combinations of possible values for a, b, n_a, n_b it is not possible to tabulate the $\hat{S}$ values. Therefore, tables 1 and 2 only provide couples (a, b) for which a success rate $\geq a/n$ is significantly higher than a success rate $\leq b/m$. The reader might have observed values of a and b that are not tabulated. Let us suppose that the observed success rate of Method A is 6/10 and the observed success rate of Method B is 1/9 (meaning that $a = 6$, $n_a = 10$, $b = 1$, $n_b = 9$). In Table 1, entry $n_a = 10$ and $n_b = 9$ contains the couple (5,1), meaning that a 5/10 success rate is significantly higher than a 1/9 success rate at 5% significance level. Since the success rate $6/10 > 5/10$ it can be deduced that Method A is significantly better than Method B (at signficance level below 5%).

n_b	n_a 2	3	4	5	6	7	8	9	10	11	12	13
2		(3,0)	(4,0)	(5,0)	(5,0)	(6,0)	(7,0)	(7,0)	(8,0)	(9,0)	(9,0)	(10,0)
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)
3				(5,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(10,1)	(11,1)	(12,1)
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)
4		(3,1)	(4,1)	(5,1)	(5,1)	(6,1)	(7,1)	(7,1)	(8,1)	(9,1)	(9,1)	(10,1)
4					(6,2)	(7,2)	(8,2)	(9,2)	(10,2)	(11,2)	(12,2)	(12,2)
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)
5		(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)
5			(4,2)	(5,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(10,2)	(10,2)	(11,2)
5							(8,3)	(9,3)	(10,3)	(11,3)	(12,3)	(13,3)
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)
6	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)
6		(3,2)	(4,2)	(5,2)	(5,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(9,2)	(10,2)
6				(5,3)	(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(12,3)
6									(10,4)	(11,4)	(12,4)	(13,4)
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)
7	(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		(3,2)	(4,2)	(4,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)
7			(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)
7					(6,4)	(7,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)
7										(11,5)	(12,5)	(13,5)
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)
8	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)
8		(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)
8		(3,3)	(4,3)	(5,3)	(5,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)
8				(5,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)
8						(7,5)	(8,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)
8												(13,6)
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)
9	(2,1)	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)
9	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)
9		(3,3)	(4,3)	(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)
9			(4,4)	(5,4)	(6,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)
9				(5,5)	(6,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(12,5)
9							(8,6)	(9,6)	(10,6)	(11,6)	(12,6)	(13,6)
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)
10	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)
10	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)
10		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)
10		(3,4)	(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)
10			(4,5)	(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)
10					(6,6)	(7,6)	(8,6)	(9,6)	(10,6)	(10,6)	(11,6)	(12,6)
10								(9,7)	(10,7)	(11,7)	(12,7)	(13,7)
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)
11	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)
11	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)
11		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(7,3)	(8,3)
11		(3,4)	(4,4)	(5,4)	(5,4)	(6,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)
11			(4,5)	(5,5)	(6,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)
11				(5,6)	(6,6)	(7,6)	(8,6)	(8,6)	(9,6)	(10,6)	(11,6)	(11,6)
11						(7,7)	(8,7)	(9,7)	(10,7)	(11,7)	(11,7)	(12,7)
11									(10,8)	(11,8)	(12,8)	(13,8)
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)
12	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)
12	(2,2)	(3,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(6,2)
12	(2,3)	(3,3)	(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)
12		(3,4)	(4,4)	(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(8,4)	(9,4)
12		(3,5)	(4,5)	(5,5)	(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)
12			(4,6)	(5,6)	(6,6)	(7,6)	(7,6)	(8,6)	(9,6)	(9,6)	(10,6)	(11,6)
12				(5,7)	(6,7)	(7,7)	(8,7)	(9,7)	(9,7)	(10,7)	(11,7)	(12,7)
12						(7,8)	(8,8)	(9,8)	(10,8)	(11,8)	(12,8)	(12,8)
12										(11,9)	(12,9)	(13,9)
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)
13	(2,1)	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(4,1)	(5,1)
13	(2,2)	(3,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)
13	(2,3)	(3,3)	(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)
13		(3,4)	(4,4)	(4,4)	(5,4)	(5,4)	(6,4)	(6,4)	(7,4)	(7,4)	(8,4)	(8,4)
13		(3,5)	(4,5)	(5,5)	(5,5)	(6,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(9,5)
13			(4,6)	(5,6)	(6,6)	(6,6)	(7,6)	(8,6)	(8,6)	(9,6)	(10,6)	(10,6)
13				(5,7)	(6,7)	(7,7)	(7,7)	(8,7)	(9,7)	(10,7)	(10,7)	(11,7)
13					(6,8)	(7,8)	(8,8)	(9,8)	(10,8)	(10,8)	(11,8)	(12,8)
13							(8,9)	(9,9)	(10,9)	(11,9)	(12,9)	(13,9)
13											(12,10)	(13,10)

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

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

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

Software and codes Codes for computing $\hat{S}$ values are publically available on the web site `http://ina.eivd.ch/projects/stamp`. There are several implementations : one in *JavaScript* that is directly interpreted by most browsers, one in *C++* and one in *Java*, thus intended for researchers that want to include the code in their own softwares.

4 Numerical results

The power of an hypothesis statistical test is defined as the probability of rejecting a false null hypothesis. So, the higher the power of an hypothesis statistical test is, the more subtle differences in the samples the test can discriminate and the better the test is considered.

This section empirically shows that the new test we propose is more powerful than McNemar ones and, for large samples, slightly more powerful than standard tests. If abusively applied to small samples, the standard test is also shown to reject a true null hypothesis with a probability higher than the significance level, showing that the standard test cannot be safely applied to small samples.

In order to show this, we proceed as follows : We choose a significance level of $\alpha = 0.01$ (which is very common in practice) and $n_a = n_b = n$ so that McNemar test could be applied. For each n, we found the lowest value of a for which one-sided McNemar test indicates that a proportion of a/n is significantly higher than a proportion of $(n-a)/n = b/n$. So, for any given n a value a is found. For both values of n and a, we find the largest value b' for which our new one-sided test indicates that a proportion of a/n is significantly (with same $\alpha = 0.01$ level) larger than a proportion of b'/n. Finally, we find the largest integer value b" for which $T = \frac{\sqrt{2n}(a-b")}{\sqrt{(a+b")(2n-a-b")}} > 2.326$, i.e. the largest value b" for which the standard test rejects the null hypothesis, (even if it is abusively applied to small n).

So, for each of the McNemar, new and standard one-tailed test, we fixed the same values of a and compared the respective values of b, b' and b", for various values of n, that are needed at most for the repective test to indicate significant different proportions.

These values are plotted on Figure 2 as a function of n. On this figure, we can see that the McNemar test is not able to discriminate proportions at $\alpha = 1\%$ level for sample sizes $n < 6$. The new test proposed in this paper is able to distinguish proportion even for samples of size 3. For the same sample size n and proportion of success a/n our new test is able to discriminate proportions b'/n much higher than the corresponding b/n proportions of McNemar test. For $n > 14$, our new test is able to discriminate proportions b'/n slightly higher than proportions $b"/n$ if a standard parametric test is applied. Finally, for $n < 7$, a standard parametric test, abusively applied, may underestimate the probability

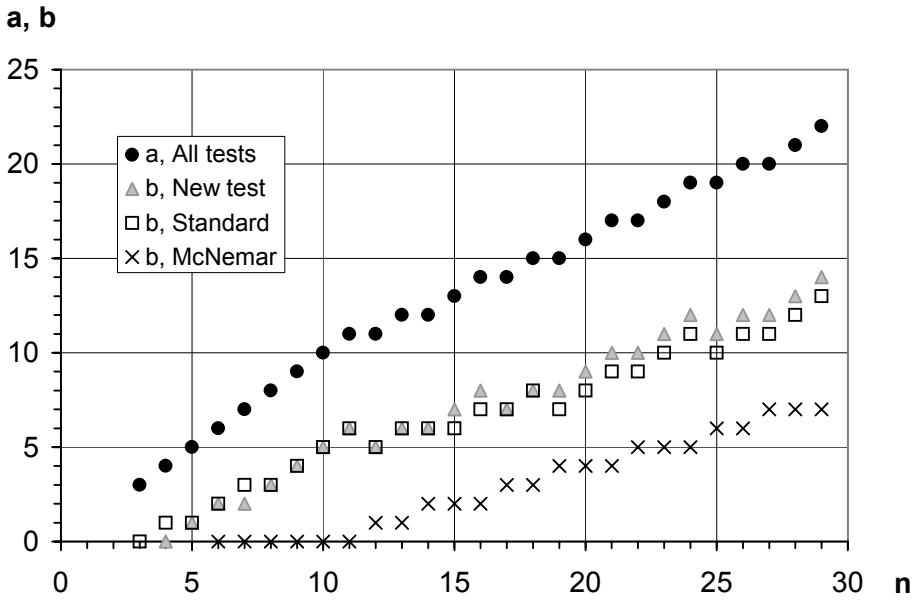

Figure 2: Values of a and b for which a proportion of a/n is shown to be significantly higher than a proportion of b/n at level $\alpha = 0.01$, for McNemar, standard and new statistical test proposed in this paper.

of occurence of the null hypothesis, if the last is true, leading to erroneous conclusions. For instance, if the null hypothesis is true and proportions of both samples is $1/2$, it is easy to show that the probability of observing $3/3$ successes for one sample and $0/3$ for the other is $1/64$, thus above 1%, whilst $T > 2.326$

Very similar figures can be drawn for various significant levels α and two-tailed tests.

5 Conclusions

The nonparametric statistical test developped in this article is shown to be much more powerful than the classical McNemar nonparametric test. The power of the nonparametric statistical test developped is comparable to standard parametric test. This result is very positive, since it is commonly believed that parametric tests are significantly more powerful than nonparametric ones. The tables provided in this article are not available in the literature and can be very useful for OR practitionners to compare proportions in a very easy way, since no computation has to be undertaken. Indeed, the user has only to count the number of positive elements in the samples to be compared.

References

[Arbuthnott(1710)] Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions*, **27**, 186–190.

[Conover(1999)] Conover, W. J. (1999). *Practical Nonparametric Statistics*. Wiley, Weinheim, 3. edition.

[Cramér(1946)] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

[Finney(1948)] Finney, D. (1948). The Fisher-Yates test of significance in 2×2 contingency tables. *Biometrika*, **35**, 145–156.

[Gail & Gart(1973)] Gail, M. H. & Gart, J. J. (1973). The determination of sample sizes for use with the exact conditional 2×2 comparative trials. *Biometrics*, **29**, 441–448.

[Garside & Mack(1976)] Garside, G. R. & Mack, C. (1976). Actual type 1 error probabilities for various tests in the homogeneity case of the 2×2 contingency table. *The American Statistician*, **30**, 18–21.

[Harkness & Katz(1964)] Harkness, W. & Katz, L. (1964). Comparison of the power functions for the test of independence in 2 x 2 contingency tables. *The Annals of Mathematical Statistics*, **35**, 1115–1127.

[McDonald et al.(1977)] McDonald, L. L., Davis, B. M., & Milliken, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables. *Technometrics*, **19**, 145–158.

[Ott & Free(1969)] Ott, R. & Free, S. (1969). A short-cut rule for a one-sided test of hypothesis for qualitative data. *Technometrics*, **11**, 197–200.

[Robertson(1960)] Robertson, W. H. (1960). Programming Fisher's exact method of comparing two percentates. *Technometrics*, **2**, 103–107.

[Taillard(1991)] Taillard, É. D. (1991). Robust taboo search for the quadratic assignment problem. *Parallel computing*, **17**, 443–455.

[Taillard & Voss(2002)] Taillard, É. D. & Voss, S. (2002). POPMUSIC: Partial OPtimization Metaheuristic Under Special Intensification Conditions. In C. Ribeiro and P. Hansen, editors, *Essays and surveys in metaheuristics*, Operations research/computer science interfaces, pages 613–629. Kluwer, Boston/Dordrecht/London.