Few statistical tests for proportions comparison

Éric D. Taillard, Philippe Waelti, Jacques Zuber

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

Abstract

This article reviews a number of statistical tests for comparing proportions. These statistical tests are presented in a comprehensive way, so that OR practitioners can easily understand them and correctly use them. A test for 2×2 contingency tables is developed and shown to be more powerful than other classical tests of the literature such as Fisher's exact test. Tables with critical values for small samples are provided, so that the test can be conducted without any computations.

keywords : Statistical test, 2×2 contingency table, method comparison

1 Introduction

In operations research, comparing two solution methods with each other is frequently needed. This is particularly the case when one wants to tune the parameters of an algorithm. In this case, one wants to know whether a given parameter setting is better than another one. In practice, to identify 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 mean, 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 mean, standard deviation, etc. of the computational effort needed to obtain the optimum solution are computed.

3. The maximal computational effort is fixed, as well as a goal to reach. One counts the number of times each method reaches the goal within the allowed computational effort.

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 only to the mean. This is evidently very insufficient for stating that a solution method is statistically better than another one.

When the standard deviation is provided in addition to the mean, it is generally (implicitly) assumed that the distribution of the population satisfies the hypothesis of a *normal distribution*. Under this assumption, a large number of statistical tests are available and can be validly performed. Unfortunately, the normality assumption is far from being always satisfied. For instance, an optimization technique that frequently finds globally optimal solutions has a distribution with a truncated tail, 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 non-deterministic optimization techniques (Robust taboo search[Taillard(1991)] and POPMUSIC[Taillard & Voss(2002)]) for a turbine runner balancing problem instance. Although this situation is frequent with metaheuristic-based optimization methods, it cannot be generalized.

This figure shows clearly that the distributions are asymmetrical, 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 mean) of an a-priori unknown distribution function is not evident. Moreover, a confidence interval for the mean should be given, which seems not evident to be undertaken. A bootstrap approach [Davison & Hinkley(2003), Efron & Tibshirani(1993)] could be convenient.

When the third comparison approach quoted above is used (counting the number of successes), the sign test[Arbuthnott(1710)] (see, e.g.[Conover(1999)]), or, better, the "Fisher's exact test" for 2×2 contingency table is convenient. A run of a method is successful if it reaches a given goal. In the context of NP-complete problems, the goal is to find a feasible solution. In the context of optimization problems (e.g. NP-hard problems), the goal could be finding the optimum solution (subject that such a solution can be characterized) or finding a solution that is a given percentage above (respectively : below) a lower (respectively : upper) bound to the optimum. When two methods have to be compared on a given set of problem instances, a success for a method could be to provide a solution of better quality than the solution produced by the other method for the same problem instance. In the context of comparing two

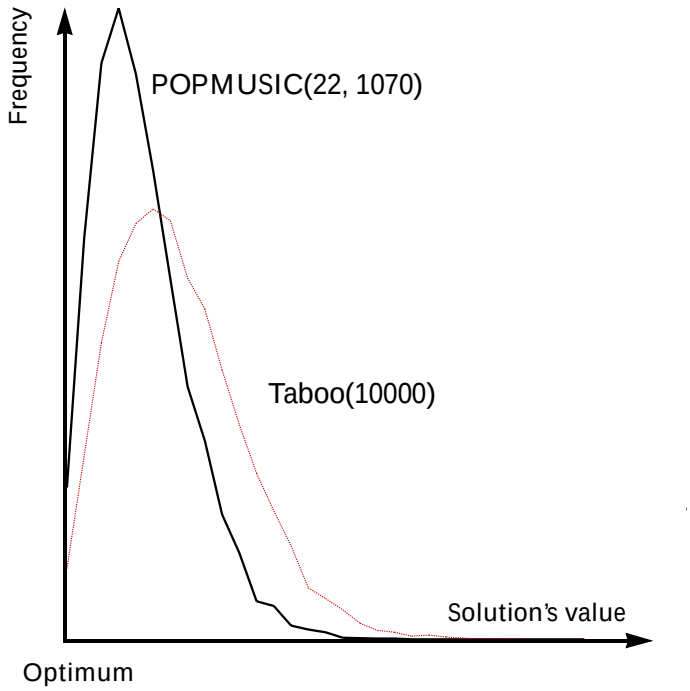

Figure 1: Empirical distributions of solution values obtained by two non-deterministic methods (POPMUSIC and taboo), obtained by solving a large number of times the same problem instance.

methods for multiobjective optimization, a success could be to find a solution that is not dominated by the set of solutions produced by the other method. Naturally, the definition of a "success" must be clearly stated before a statistical test is undertaken, but the user has a wide latitude in choosing the definition, possibly leading to different conclusions !

This article develops a statistical test that is more powerful than both the sign test and Fisher's test for comparing proportions. This test is based on a standard methodology that seems to be uncommon in practice. Indeed, it is not developed in the literature consulted, although it cannot be excluded that it appears somewhere, since there is a huge amount of articles and books dealing with contingency tables (see, e.g.[Conover(1999), Good(2005)]). Before the presentation of the new test, other approaches commonly used in practice are reviewed. Finally, the new test is numerically compared to these approaches.

2 Comparing proportions

The central problem treated by this article is the following : Let us suppose that two populations A and B are governed by Bernoulli distributions, i.e., the probability of success of an occurrence 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 annealing, or the problem data can be viewed as random, the user not being 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 successful.

Say that the user would like to use the method having the highest success probability. Ideally, the user would like to know p_a and p_b to make the choice. Unfortunately, these probabilities are unknown. The user could try to estimate them empirically. In the remaining of the paper, we assume the following :

Assumptions and sampling

- The sample size of A is n_a ; a successes and $n_a - a$ failures have been observed.
- The sample size of 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) of having a success for population A (respectively : B) does not depends on the observations. Either $p_a < p_b$ or $p_b < p_a$ or $p_a = p_b$ for all observations (unbiased sample).

	Success	Failure	Total
Sample 1	a	$n_a - a$	n_a
Sample 2	b	$n_b - b$	n_b
Total	$a + b$	$n_a + n_b - a - b$	$n_a + n_b$

Table 1: 2×2 Contingency table

Hence, the data can be put in a 2×2 contingency table, as shown in Table 2.

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 mean 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 quoted without proper justification is $\min(n_a \cdot p_a \cdot q_a, n_b \cdot p_b \cdot q_b) > 5$), one can approximate D by a normal distribution.

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

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

Alternative hypothesis (one-sided test) $p_a > p_b$.

We very briefly recall here the principle of a hypothesis statistical test. First, a *null hypothesis* — that the user would like to be rejected by the test — is chosen, as well as the logical negation of the null hypothesis, the *alternate hypothesis*. Then, a test statistic T is chosen for which the distribution is known if the null hypothesis is true. A series of n experiments is then performed, whose results allow to compute the value of the considered test statistic. With the help of the distribution of T, one computes the probability (known as p-value) of observing the value obtained by the experiments, if the null hypothesis is true. The smaller the p-value, the more reasonable it is to reject the null hypothesis. Generally, before proceeding to the experiments, a significance level α is chosen, typically 0.05 or 0.01. The null hypothesis is rejected if the p-value is lower than α.

In conducting a hypothesis statistical test, the null hypothesis is chosen in such a way that it is felt not to be true. So, in the above mentioned 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 obtained by inverting the roles of A and B. Since probabilities p_a and p_b are unknown, estimators $\hat{p}_a$ and $\hat{p}_b$ that would maximize the probability of null hypothesis to be true are sought. This maximum occurs for $\hat{p}_a = \hat{p}_b = \hat{p}$. The pooled estimate : $\hat{p} = \frac{a+b}{n_a+n_b}$ is the best estimate of the common value of the probability of success [Barnes(1994)]. The value observed for d is estimated by $\hat{d} = a/n_a - b/n_b$ and the variance σ_D^2 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 $\hat{d}/\hat{s} > 1.645 = \Phi^{-1}(1 - 0.05)$
- $\alpha = 1\%$ if $\hat{d}/\hat{s} > 2.326$
- $\alpha = 0.1\%$ if $\hat{d}/\hat{s} > 3.09$

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

Null hypothesis $p_a = p_b$

Alternative hypothesis (two-sided test) $p_a \neq p_b$

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

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

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 not excluded) at significance levels :

- $\alpha = 5\%$ if $T > 3.841 = \chi_{1;1-0.05}$
- $\alpha = 1\%$ 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 significance of changes

The sign test is perhaps the first nonparametric test ever published [Arbuthnott(1710)]. A variation of this test is known as McNemar test for significance of changes.

In many situations, both samples are of the 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 whether Method A is significantly 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 (as soon as the binomial distribution can be computed), since it is not based on the central limit theorem.

2.3 Nonparametric Fisher's exact test

The Fisher's exact test is a *permutation* test based on the following idea. Suppose that a successes are observed for a sample (of size n_a) from a first popula-

tion and b successes are observed for sample (of size n_b) from a second population. Assume that the proportion of successes is the same for both populations (null hypothesis) and that the marginals of the 2×2 contingency table are fixed ("Total" line and column of Table 2). Under the null hypothesis, the successes can be distributed independently of one another among the first or second samples, provided that the marginals are kept identical to the observed values. There is a total of $C_{a+b}^{n_a+n_b}$ different ways to distribute the successes (or, equivalently, to build 2×2 contingency tables with same marginals). The one-tailed Fisher's exact test simply counts the number of contingency tables with the same marginals that are as extreme as or more extreme than the original table. The p-value of one-tailed Fisher's exact test is the ratio of the number of tables at least as extreme as the original table over the total number of tables with the same marginals.

There are several variants for the two-tailed Fisher's test. One that is commonly implemented is to count the number of tables with the same marginals that have a probability not higher than the original table to occur under the null hypothesis. The two-tailed Fisher's test is the ratio of this number over the total number of possible table with the same marginals. This variant is known as "Tocher's modification" of Fisher's test [Siegel(1956)] and will be used in the numerical results that follows.

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 successful 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 not be significant, even if Method B was always successful ($b = n_b$), whilst he could choose a larger value for n_a (and thus get significant differences). Therefore, we developed a new statistical test for comparing proportions. We have observed that this new test is more powerful than Fisher's one.

3.1 One-sided test

For the one-sided test, let us suppose that it is known that p_a cannot be lower than p_b. Note that this assumption must also be done for a one-sided Fisher's test. This situation arises in OR when a second method cannot produce worse

results than a first one, e.g. the second method explores a superset of solutions examined by the first method, or the second method tries to improve the solutions produced by the first one. Hence, it is assumed that the user has observed $a/n_a > b/n_b$ and wants to show that population A has a strictly higher success rate than population B.

Null hypothesis $p_a = p_b = p$

Alternative hypothesis $p_a > p_b$

The methodology applied in our test is quite standard and consists of directly computing the probability $S(p)$ to observe a successes or more with n_a observations and b successes or less with n_b observations, under null hypothesis. This probability is given by :

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

This probability depends on proportion p which is unknown. Since the null hypothesis is to be rejected with the highest security, probability $S(p)$ must be maximized over p.

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

$$\hat{S} = \max_{0<p<1} S(p) < \alpha$$

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)]), but the last is fully nonparametric. In our developments we make use of parameter p. Note however that if our new test concludes that the null hypothesis should be rejected, this conclusion remains valid whatever the common value of p is.

3.1.1 Examples

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

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

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

.

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 $S(p)$. For instance, if $n_a = 3$ and $n_b = 2$, $S(\frac{3+0}{3+2}) = \hat{S} = 108/3125 < 5\%$. So a success rate of $3/3$ is significantly higher than a success rate of $0/2$, if the decision rule is to reject p-values lower than $\alpha = 0.05$.

Unfortunately, for arbitrary values of a, n_a, b and n_b, the pooled estimate is *not* the value that maximizes $S(p)$. For instance, for $a = 3$, $n_a = 4$, $b = 0$ and $n_b = 3$, $S(p) = C_3^4 p^3(1-p)^4 + C_4^4 p^4(1-p)^3$. So, $S(3/7) < 4/100$ while $S(\frac{6-2\sqrt{2}}{7}) > 4/100$.

This means that if the significance level is fixed at $\alpha = 0.04$, an erroneous conclusion will be drawn if the pool estimate is used for testing if a rate of $3/4$ is significantly higher than a rate of $0/3$.

Although the difference in $S(p)$ values for the above example is not very large, the pooled estimate may underestimate by more than $1/3$ the $\hat{S}$ value. This is exemplified by a success rate of $4/4$ compared to a success rate of $56/100$. The pooled estimate would provide $S(60/104) < 4.5\%$ while there is a value of p that provides a value of $\hat{S}$ near to 6%.

3.1.2 Computation of decision rule

In general, the analytical expression of $\hat{S}$ is hard to find in practice. Therefore, we have numerically estimated $\hat{S}$. We provide in Table 2 (and, respectively, in Table 3), 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 possible combinations of values for a, b, n_a and n_b it is not possible to tabulate the $\hat{S}$ values. Therefore, tables 2 and 3 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 2, entry $n_a = 10$ and $n_b = 9$ contains the couple (5,1), meaning that a $5/10$ success rate can be considered as significantly higher than a $1/9$ success rate at 5% significance level. Since $6/10 > 5/10$, it can be deduced that Method A is significantly better than Method B (at significance level below 5%).

Conversely, if the significance level is 1%, we can see in Table 3 that the couple $(7, 1)$ is contained in the entry $n_a = 10$ and $n_b = 9$. So, we cannot reject that a success rate of $6/10$ is equal to a success rate of $1/9$, at significance level

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 2: 5 % One-tailed test. Couples (a, b) for which a success rate $\geq a/n_a$ can be considered as higher than a success rate $\leq b/n_b$, for a 5 % significance level. Couples in boldface indicate that Fisher's unilateral test has a p-value strictly higher than 0.05.

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 3: 1 % One-tailed test. Couples (a, b) for which a success rate $\geq a/n_a$ can be considered as higher than a success rate $\leq b/n_b$, for a 1% significance level. Couples in boldface indicate that Fisher's unilateral test has a p-value strictly higher than 0.01

of 1%.

3.2 Two-sided test

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

Null hypothesis $p_a = p_b = p$

Alternative hypothesis $p_a \neq p_b$

To simplify the developments that follow, we suppose that $a/n_a > b/n_b$. This is not limitative since both samples can be permuted. In order to conduct a two-sided test, it can be considered that two one-sided tests must be simultaneously conducted, — the first one with alternative hypothesis $p_a > p_b$ and a second one with alternative hypothesis $p_a < p_b$. Basically, the one-sided test developed above provides a p-value. For the two-sided text, both p-values can be added. Therefore, computing the probability $T(p)$ of occurrence of the observations under null hypothesis can be decomposed into two parts. The first part takes into account the summations $\sum_{i=a}^{n_a} \sum_{j=0}^{b}$ and the symmetrical part with summations $\sum_{i=0}^{n_a-a} \sum_{j=n_b-b}^{n_b}$. In order to get a good estimate $\hat{T}$ for the maximum of $T(p)$ (better than $2\hat{S}$), both summations must be done with a common proportion value p. So, the following decision rule can be formulated :

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

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

This decision rule is somewhat harder to compute than for the one-tailed test. Indeed, $T(p)$ can have different optima when varying the common proportion p. Without considering the extreme cases ($a = 0$ and $b = 0$, or $a = n_a$ and $b = n_b$) for which the null hypothesis cannot be rejected, the distribution is symmetrical with $p = 0.5$ being an optimum value — either the global maximum or a local *minimum*. The last case occurs when the ratios a/n_a and b/n_b are relatively different, precisely when the test is interesting to be conducted ! So, we have numerically estimated $\hat{T}$ and provide in Table 4 (and, respectively, in Table 5) — 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 equal to a b/n_b rate.

n_b \ n_a	2	3	4	5	6	7	8	9	10	11	12	13
2			**(4, 0)**	(5,0)	(6,0)	**(6, 0)**	**(7, 0)**	**(7, 0)**	**(8, 0)**	**(9, 0)**	**(9, 0)**	**(10, 0)**
3		(3,0)	(4,0)	**(4, 0)**	(5,0)	(6,0)	**(6, 0)**	(7,0)	**(7, 0)**	**(8, 0)**	**(8, 0)**	**(9, 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)**	(4,0)	**(4, 0)**	**(5, 0)**	**(5, 0)**	**(6, 0)**	**(6, 0)**	**(7, 0)**	**(7, 0)**	**(8, 0)**
4			**(4, 1)**	(5,1)	(6,1)	**(6, 1)**	**(7, 1)**	**(8, 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)**	**(5, 0)**	**(5, 0)**	**(5, 0)**	**(6, 0)**	**(6, 0)**
5		**(3, 1)**	(4,1)	(5,1)	**(5, 1)**	**(6, 1)**	(7,1)	**(7, 1)**	**(8, 1)**	**(8, 1)**	**(9, 1)**	(10,1)
5				**(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		(3,1)	(4,1)	**(4, 1)**	(5,1)	**(5, 1)**	**(6, 1)**	(7,1)	**(7, 1)**	(8,1)	**(8, 1)**	**(9, 1)**
6			**(4, 2)**	**(5, 2)**	(6,2)	**(6, 2)**	**(7, 2)**	**(8, 2)**	**(8, 2)**	**(9, 2)**	**(10, 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)**	**(5, 1)**	**(5, 1)**	**(6, 1)**	**(6, 1)**	**(7, 1)**	**(7, 1)**	**(8, 1)**
7			**(4, 2)**	(5,2)	**(5, 2)**	**(6, 2)**	(7,2)	**(7, 2)**	**(8, 2)**	(9,2)	**(9, 2)**	**(10, 2)**
7				**(5, 3)**	**(6, 3)**	(7,3)	**(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)**	**(7, 1)**
8		**(3, 2)**	(4,2)	(5,2)	**(5, 2)**	(6,2)	**(6, 2)**	**(7, 2)**	**(7, 2)**	**(8, 2)**	**(9, 2)**	**(9, 2)**
8			**(4, 3)**	**(5, 3)**	(6,3)	**(6, 3)**	**(7, 3)**	(8,3)	**(8, 3)**	**(9, 3)**	**(10, 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)	(4,2)	**(4, 2)**	(5,2)	**(5, 2)**	**(6, 2)**	**(6, 2)**	**(7, 2)**	**(7, 2)**	**(8, 2)**	(9,2)
9			**(4, 3)**	(5,3)	(6,3)	**(6, 3)**	(7,3)	**(7, 3)**	**(8, 3)**	**(9, 3)**	**(9, 3)**	**(10, 3)**
9				**(5, 4)**	(6,4)	(7,4)	**(7, 4)**	**(8, 4)**	**(9, 4)**	**(9, 4)**	**(10, 4)**	**(11, 4)**
9					**(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)**	**(6, 2)**	**(6, 2)**	**(7, 2)**	**(7, 2)**	**(8, 2)**
10		**(3, 3)**	(4,3)	(5,3)	**(5, 3)**	(6,3)	**(6, 3)**	**(7, 3)**	(8,3)	**(8, 3)**	**(9, 3)**	**(9, 3)**
10			**(4, 4)**	(5,4)	(6,4)	**(6, 4)**	**(7, 4)**	**(8, 4)**	**(8, 4)**	**(9, 4)**	**(10, 4)**	**(10, 4)**
10				**(5, 5)**	**(6, 5)**	(7,5)	**(7, 5)**	**(8, 5)**	**(9, 5)**	**(10, 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)**	**(6, 2)**	**(6, 2)**	**(6, 2)**	**(7, 2)**
11		**(3, 3)**	(4,3)	**(4, 3)**	(5,3)	(6,3)	**(6, 3)**	(7,3)	**(7, 3)**	(8,3)	**(8, 3)**	**(9, 3)**
11			**(4, 4)**	(5,4)	(6,4)	**(6, 4)**	**(7, 4)**	**(7, 4)**	**(8, 4)**	(9,4)	**(9, 4)**	**(10, 4)**
11				**(5, 5)**	(6,5)	(7,5)	**(7, 5)**	**(8, 5)**	**(9, 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)**	(5,3)	**(5, 3)**	**(6, 3)**	**(6, 3)**	**(7, 3)**	**(7, 3)**	**(8, 3)**	**(8, 3)**
12		**(3, 4)**	(4,4)	(5,4)	**(5, 4)**	**(6, 4)**	(7,4)	**(7, 4)**	(8,4)	**(8, 4)**	**(9, 4)**	(10,4)
12			**(4, 5)**	(5,5)	(6,5)	**(6, 5)**	**(7, 5)**	**(8, 5)**	**(8, 5)**	**(9, 5)**	(10,5)	**(10, 5)**
12				**(5, 6)**	**(6, 6)**	(7,6)	**(7, 6)**	**(8, 6)**	**(9, 6)**	**(9, 6)**	**(10, 6)**	**(11, 6)**
12					**(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)**	**(6, 3)**	**(6, 3)**	**(6, 3)**	**(7, 3)**	**(7, 3)**
13		**(3, 4)**	(4,4)	(5,4)	**(5, 4)**	(6,4)	**(6, 4)**	(7,4)	**(7, 4)**	**(8, 4)**	(9,4)	**(9, 4)**
13			**(4, 5)**	(5,5)	(6,5)	**(6, 5)**	**(7, 5)**	(8,5)	**(8, 5)**	(9,5)	**(9, 5)**	**(10, 5)**
13				**(5, 6)**	(6,6)	(7,6)	**(7, 6)**	**(8, 6)**	**(9, 6)**	**(9, 6)**	**(10, 6)**	**(10, 6)**
13				**(5, 7)**	**(6, 7)**	**(7, 7)**	(8,7)	**(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 4: 5 % Two-tailed test. Couples (a, b) for which a success rate $\geq a/n_a$ can be considered different than a success rate $\leq b/n_b$, for 5 % significance level. Couples in boldface indicate that Fisher's bilateral test has a p-value strictly higher than 0.05.

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				**(5, 0)**	**(6, 0)**	(7,0)	**(7, 0)**	**(8, 0)**	**(9, 0)**	**(9, 0)**	**(10, 0)**	**(11, 0)**
3											**(12, 1)**	**(13, 1)**
4			**(4, 0)**	(5,0)	(6,0)	**(6, 0)**	**(7, 0)**	(8,0)	**(8, 0)**	**(9, 0)**	**(9, 0)**	**(10, 0)**
4						**(7, 1)**	**(8, 1)**	**(9, 1)**	**(10, 1)**	(11,1)	**(11, 1)**	**(12, 1)**
5		**(3, 0)**	(4,0)	(5,0)	**(5, 0)**	(6,0)	**(6, 0)**	**(7, 0)**	(8,0)	**(8, 0)**	(9,0)	**(9, 0)**
5					**(6, 1)**	**(7, 1)**	(8,1)	**(8, 1)**	**(9, 1)**	**(10, 1)**	**(10, 1)**	(11, 1)
5								**(9, 2)**	**(10, 2)**	**(11, 2)**	**(12, 2)**	**(13, 2)**
6		**(3, 0)**	(4,0)	**(4, 0)**	(5,0)	**(5, 0)**	(6,0)	**(6, 0)**	**(7, 0)**	**(7, 0)**	**(8, 0)**	**(8, 0)**
6				**(5, 1)**	(6,1)	(7,1)	**(7, 1)**	**(8, 1)**	(9,1)	**(9, 1)**	**(10, 1)**	(11,1)
6						**(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)	(6,1)	**(6, 1)**	**(7, 1)**	(8,1)	**(8, 1)**	**(9, 1)**	**(9, 1)**	**(10, 1)**
7					**(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)**	(5,1)	**(5, 1)**	**(6, 1)**	(7,1)	**(7, 1)**	(8,1)	**(8, 1)**	**(9, 1)**	(9, 1)
8				**(5, 2)**	(6,2)	(7,2)	(8,2)	**(8, 2)**	**(9, 2)**	**(9, 2)**	**(10, 2)**	**(11, 2)**
8						**(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)**	(8,1)	**(8, 1)**	**(9, 1)**
9				**(5, 2)**	(6,2)	(7,2)	**(7, 2)**	(8,2)	**(8, 2)**	**(9, 2)**	(10,2)	**(10, 2)**
9					**(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)	(6,2)	**(6, 2)**	(7,2)	**(7, 2)**	**(8, 2)**	(9,2)	**(9, 2)**	**(10, 2)**
10					**(6, 3)**	(7,3)	(8,3)	**(8, 3)**	**(9, 3)**	(10,3)	**(10, 3)**	**(11, 3)**
10						**(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)**	(8,2)	**(8, 2)**	**(9, 2)**	**(9, 2)**
11				**(5, 3)**	(6,3)	(7,3)	**(7, 3)**	(8,3)	(9,3)	**(9, 3)**	**(10, 3)**	**(11, 3)**
11					**(6, 4)**	**(7, 4)**	**(8, 4)**	(9,4)	**(9, 4)**	**(10, 4)**	(11,4)	**(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)**	**(8, 2)**	**(8, 2)**	**(9, 2)**
12			**(4, 3)**	(5,3)	(6,3)	**(6, 3)**	**(7, 3)**	(8,3)	**(8, 3)**	**(9, 3)**	(10,3)	**(10, 3)**
12					**(6, 4)**	**(7, 4)**	(8,4)	**(8, 4)**	**(9, 4)**	(10,4)	**(10, 4)**	**(11, 4)**
12						**(7, 5)**	**(8, 5)**	(9,5)	**(9, 5)**	**(10, 5)**	**(11, 5)**	**(12, 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)**	**(8, 2)**	**(8, 2)**
13			**(4, 3)**	(5,3)	(6,3)	**(6, 3)**	(7,3)	**(7, 3)**	**(8, 3)**	(9,3)	**(9, 3)**	(10,3)
13				**(5, 4)**	**(6, 4)**	(7,4)	**(7, 4)**	**(8, 4)**	(9,4)	**(9, 4)**	**(10, 4)**	(11,4)
13					**(6, 5)**	**(7, 5)**	(8,5)	(9,5)	**(9, 5)**	**(10, 5)**	**(11, 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)**	**(9, 3)**	**(9, 3)**
14				**(5, 4)**	(6,4)	(7,4)	**(7, 4)**	(8,4)	**(8, 4)**	**(9, 4)**	(10,4)	**(10, 4)**
14					**(6, 5)**	(7,5)	(8,5)	**(8, 5)**	**(9, 5)**	**(10, 5)**	**(10, 5)**	**(11, 5)**
14						**(7, 6)**	**(8, 6)**	(9,6)	**(9, 6)**	**(10, 6)**	**(11, 6)**	**(11, 6)**
14							**(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)**	**(8, 3)**	**(8, 3)**	**(9, 3)**
15			**(4, 4)**	(5,4)	(6,4)	**(6, 4)**	(7,4)	(8,4)	**(8, 4)**	(9,4)	**(9, 4)**	**(10, 4)**
15				**(5, 5)**	**(6, 5)**	(7,5)	(8,5)	**(8, 5)**	**(9, 5)**	**(9, 5)**	**(10, 5)**	(11,5)
15					**(6, 6)**	**(7, 6)**	(8,6)	(9,6)	**(9, 6)**	**(10, 6)**	**(11, 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 5: 1 % Two-tailed test. Couples (a, b) for which a success rate $\geq a/n_a$ can be considered different than a success rate $\leq b/n_b$, for 1% significance level. Couples in boldface indicate that Fisher's bilateral test has a p-value strictly higher than 0.01.

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

4 Numerical results

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

This section empirically shows that the new test we propose is more powerful than those provided by McNemar and Fisher 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 find the lowest value of a for which a 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 larger than a proportion of b'/n (with same $\alpha = 0.01$ level). Then, we find (manually) 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). Finally, we find the largest values b''' for which Fisher's exact test rejects the null hypothesis.

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

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 proportions even for samples of size 3. For the same sample size n and a 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 and slightly higher than b'''/n proportions of Fisher's 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 of occurrence of the null hypothesis, 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 significance levels α and two-tailed tests.

5 Conclusions

The new statistical test developed in this article is shown to be much more powerful than the classical McNemar nonparametric test. The power of the new statistical test developed is comparable to standard parametric test and slightly more powerful than Fisher's exact test. This result is very positive, since it is commonly believed that parametric tests are significantly more powerful than nonparametric ones and that Fisher's exact test is the best for 2×2 contingency tables. The tables provided in this article are not available in the literature and can be very useful to OR practitioners to compare proportions in a very easy way, since no computation has to be undertaken. Indeed, the user only has to count the number of positive elements in the samples to be compared.

6 Acknowledgements

The authors are grateful to anonymous referees whose comments help in improving this article. They would like to thank C. Evquoz for corrections brought to the manuscript. This research was partially supported by the University of Applied Sciences of Western Switzerland, grant QUALOPT-11731.

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.

[Barnes(1994)] Barnes, J. W. (1994). *Statistical Analysis for Engineers and Scientists*. McGraw-Hill, New-York.

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

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

[Davison & Hinkley(2003)] Davison, A. C. & Hinkley, D. (2003). *Bootstrap Methods and their Application*. Cambridge University Press, 5th edition.

[Efron & Tibshirani(1993)] Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman-Hall.

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

[Garside & Mack(1976)] Garside, G. & 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**(1), 18–21.

[Good(2005)] Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypothesis*. Springer, New-York, third edition.

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

[McDonald et al.(1977)] McDonald, L., Davis, B., & Milliken, G. (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. (1960). Programming Fisher's exact method of comparing two percentages. *Technometrics*, **2**, 103–107.

[Siegel(1956)] Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New-York.

[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). *Essays and surveys in metaheuristics (C. Ribeiro, P. Hansen, eds)*, chapter POPMUSIC: Partial Optimization Metaheuristic Under Special Intensification Conditions, pages 613–629. Operations research/computer science interfaces. Kluwer, Boston/Dordrecht/London.

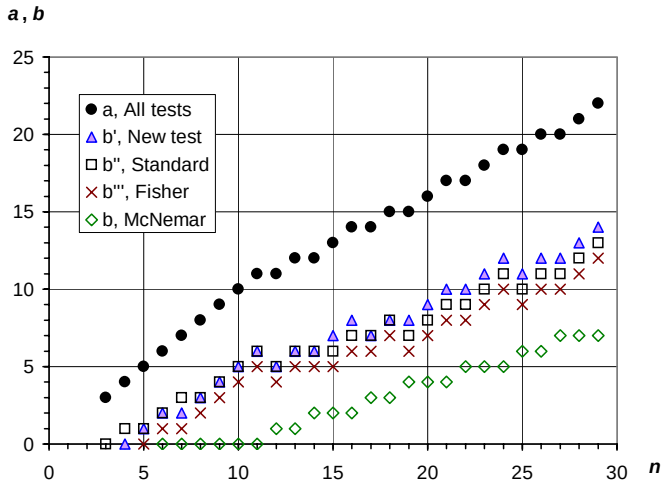

Figure 2: Values of a and b for which a proportion of a/n is shown to be significantly higher than a proportion of b/n at level $\alpha = 0.01$, for McNemar, standard, Fisher and new statistical test proposed in this paper.