Gest. Prod., São Carlos, v. 23, n. 3, p. 588-599, 2016
http://dx.doi.org/10.1590/0104-530X357-15

Non-parametric tests for small samples of categorized variables: a study

Testes não paramétricos para pequenas amostras de variáveis não categorizadas: um estudo

José Luiz Contador[1]
Edson Luiz França Senne[2]

Abstract: This paper presents a study on non-parametric tests to verify the similarity between two small samples of variables classified into multiple categories. The study shows that the only tests available for this situation are the chi-square and the exact tests. However, asymptotic tests, such as the chi-square, may not work well for small samples, leaving exact tests as the alternative. Nevertheless, if the number of classes increases, the implementation of these tests can become very difficult, in addition to requiring specific algorithms that may demand considerable computational effort. Therefore, as an alternative to the exact tests, a new test based on the difference between two uniform distributions is proposed. Computational assays are conducted to evaluate the performance of these three tests. Although non-parametric tests present numerous applications in various areas of knowledge, this study was motivated by the need to verify whether the business strategy adopted by a company is a determining factor for its competitiveness.

Keywords: Non-parametric tests; Small samples; Computer simulation; Competitive strategy.

Resumo: *Apresenta-se neste trabalho um estudo sobre testes não paramétricos para verificar a semelhança entre duas pequenas amostras de variáveis classificadas em múltiplas categorias. Mostra-se que, para essa situação, os únicos testes disponíveis são qui-quadrado e os testes exatos. Porém, testes assintóticos (como o qui-quadrado) podem não funcionar bem para pequenas amostras, sobrando como alterativa a aplicação de testes exatos. Mas, se o número de categorias cresce, a aplicação desses testes pode-se tornar bastante difícil, além de requerer algoritmos específicos, que podem exigir grande esforço computacional. Assim, um novo teste baseado na diferença de duas distribuições uniformes é proposto como uma alternativa ao teste exato. Ensaios computacionais são realizados para avaliar o desempenho desses três testes. Embora testes não paramétricos tenham inúmeras aplicações em diversas áreas de conhecimento, este trabalho surgiu motivado pela necessidade de verificar se a estratégia de negócio adotada pela empresa é um fator determinante para sua competitividade.*

Palavras-chave: *Testes não paramétricos; Pequenas amostras; Simulação computacional; Estratégia competitiva.*

1 Introduction

This work was motivated by a need to create an easily applied statistical test to aid research based on the development of the Fields and Weapons of Competition (*FWC*) model (Contador, 2008) to gauge (among other things) whether the business strategy adopted by a company is a determining factor of its competitiveness. In his research, the author of this model collected a small sample of companies and divided them into two groups. One group was formed by the most competitive companies and the other by the least competitive. The test is used to determine whether both groups adopt similar business strategies (null hypothesis H_0).

The proposed test can be used for any problem with the following characteristics:

a) Two different groups, *I* and *II* (for example, more competitive and less competitive companies), representing samples of larger populations, with n_1 and n_2 elements in each group, where n_1 and n_2 are small values;

[1] Programa de Pós-graduação em Administração das Micro e Pequenas Empresas, Faculdade Campo Limpo Paulista – FACCAMP, Rua Guatemala, 167, Jardim América, CEP 13231-230, Campo Limpo Paulista, SP, Brazil, e-mail: jluiz@feg.unesp.br

[2] Faculdade de Engenharia, Universidade Estadual Paulista – UNESP, Campus de Guaratinguetá, Av. Ariberto Pereira da Cunha, 333, CEP 12516-410, Guaratinguetá, SP, Brazil, e-mail: elfsenne@feg.unesp.br

Received Oct. 2, 2014 - Accepted Dec. 18, 2015
Financial support: The first author acknowledges financial support from CNPq (DT 307363/2015-5). The second author acknowledges financial support from CNPq (grant 303339/2013-6).

b) For each group or sample, the random variable assumes values of frequencies in each of the m classes, $m>2$ (see Table 1), i.e., the random variable is measured on a nominal scale or categorized with more than two categories

c) The number of classes or categories that the random variable may assume (value of m) is moderate in relation to the n_1 and n_2 values

It should be noted that if the random variable could be classified into only two categories (e.g., two strategies), the problem could be easily solved by Fisher's exact test (see Section 4), whatever the size of n_1 and n_2 of the samples from the two groups.

If, on the other hand, there were more than two categories for the random variable, but for each class a sufficiently large number of individuals (which would generate a problem with large samples), it would also be easy to determine the similarity between the two sets of responses using the chi-square text, which can fail when small samples are involved.

The other non-parametric tests that are available (sign test, Wilcoxon signal rank test, rank sum test, median test and t-test for paired dataset) are inadequate, as will be demonstrated through examples. Thus, for the case of small samples and more than two classes for the random variable, the problem is difficult to solve.

Therefore, the only safe alternative for addressing this type of problem is exact tests, such as the one presented in StatXact (2008), with the solution based on an extension of Fisher's Exact Test (Fisher, 1970) proposed by Freeman & Halton (1951). However, the implementation of this test requires specific algorithms and, in some cases, requires considerable computational effort, which justifies the search for new tests for this type of problem.

In light of this, the present article presents a comparative performance study (capacity to decide H_0 correctly) of the exact tests, chi-square and a new test based on the difference between two uniform distributions, proposed here. The effectiveness of these tests is compared using three indicators (risks α and β and the characteristic indicator, *CI*, extracted from the power curve, which will be constructed through simulation.

Table 1. Frequencies of strategies (*FC*) for the groups of companies.

FC	j	f_j	g_j
A	1	2	4
B	2	1	0
C	3	3	2
D	4	2	1
E	5	2	2
F	6	0	2

The studies developed here focus on attempting to solve the problem of strategy related to the *FWC* model. For this reason, some concepts of this model are given in the following section, as they are essential for understanding the problem in question. The aim of this article is not to discuss or introduce the *FWC* model. If the reader would like to know more about the model, a source of further information is provided in the references.

Numerous other problems related to biology, medicine and the social and human sciences have the characteristics described above and could be addressed using the statistical techniques used here. Some examples of problems directly related to social engineering are:

- Determining whether two different types of employees (machine operators and office workers, for example) in small companies (with few workers) have similar motivations in order to develop a single incentives program (or include all workers in a single program);
- Determining, through a small sample of companies from different sectors (e.g., manufacturing and services) whether these companies value the same characteristics in their executives to standardize human development programs;
- Determining whether executives (few in number) from different business units of a corporation have similar managerial capacity;
- Determining whether two different production processes, by analyzing few parts, create products with similar levels of quality for different characteristics (size, finishing, etc.).

The main result of the work was that effectiveness of the proposed test was similar to that of exact tests and that it performs well in situations in which the chi-square test fails (small samples and scanty, unbalanced data). Therefore, it is a real alternative to the exact test, the application of which often requires special software with restricted access.

In Section 3, there is a brief discussion on non-parametric tests and a critical analysis of their application to solve the problem in question (strategy). In Section 4, the solution adopted by the *StatXact* for problems with categorized variables is presented. In Section 5, the development of the proposed test is presented, based on the difference between two uniform distributions. In Section 6, the studies conducted to assess the performance of the

three tests (the proposed test, the exact test and the chi-square) are presented. The conclusions are given in Section 7. This final section also shows how the proposed test can be extended for problems with more than two independent samples, and are presented two examples in which the proposed test shows a clear advantage over the chi-square.

2 Fields and weapons of competition model

According to the *FWC* model, companies focus their competitive strategy on one of the 14 fields of competition (clustered in five macro fields), although they can adopt another (two or three) supporting fields. The fields of competition, according to the *FWC* model, are as follows:

- *Macro-field of competition in price*: (1) the price itself, (2) payment conditions, and (3) prize and/or promotion;
- *Macro-field of competition in product, goods or services*: (4) product project, (5) product quality, and (6) variety of models;
- *Macro-field of competition in attendance*: (7) presales technological service, (8) assistance during sale, and (9) after-sales technical service;
- *Macro-field of competition in delivery time*: (10) deadline of budgeting and negotiation, and (11) product delivery deadline;
- *Macro-field of competition in image*: (12) product and brand name, (13) reliability of the company, and (14) social responsibility (civil and preservationist).

The test of the *FWC* model assumes that a company's competitiveness is not determined by its choice of competitive strategy. Rather, it is the correct alignment of its core competence (Hamel & Prahalad, 1995) with the chosen field of competition, whatever it may be. Evidently, the model assumes that it is necessary to choose for each product/market pair one of the fields that is of interest to the market.

For a better understanding of the problem in question, consider the data in Chart 1, extracted from one of the studies conducted by Contador (2008), with the set of 21 companies which, by degree of competitiveness (*DC*), were divided into two groups: the most and least competitive. To determine the degree of competitive of the company *i* (DC_i), the *FWC* model normally uses the variation that occurs in a given period of time for invoicing or net turnover of the company.

A company *i* is classified as being in the group of the most or least competitive, in the *FWC* model, using the Nihans index (*N*). For a group of *n* companies, the Nihans index is calculated using the following formula (Equation 1):

$$N = \frac{\sum_{i=1}^{n} GC_i^2}{\sum_{i=1}^{n} GC_i} \quad (1)$$

Thus, if $DC_i \geq N$, then the company is classified among the most competitive. Otherwise, it is considered as among the least competitive.

The *FC* column for each group of companies in Chart 1 represents the codes of the main fields of competition declared by the respective companies. Thus, the strategies of both groups of companies can be represented by the C_1 lists (Set 1 - the most

Chart 1. Classification of companies in the most and least competitive groups.

Group I: Most competitive companies				Group II: Least competitive companies			
Code	Main field of competition (*FC*) Denomination	*FC*	DC_i	Code	Main field of competition (*FC*) Denomination	*FC*	DC_i
E10	Product and brand image	A	1.51	E05	Variety of models	D	0.82
E13	Product delivery deadline	B	1.43	E11	After-sales service	C	0.80
E17	After-sales service	C	1.39	E06	Product and brand image	A	0.79
E19	After-sales service	C	1.32	E12	Product and brand image	A	0.79
E21	Variety of models	D	1.25	E04	Product and brand image	A	0.69
E02	Product and brand image	A	1.19	E14	Presales service	F	0.62
E08	Product project	E	1.16	E16	Product project	E	0.54
E03	After-sales service	C	1.14	E07	Product and brand image	A	0.47
E13	Product project	E	1.11	E09	Presales service	F	0.38
E01	Variety of models	D	1.07	E20	Product project	E	0.30
				E18	After-sales service	C	0.25

Source: Contador (2008).

competitive companies) and the C_2 (Set 2 - the least competitive companies), as follows:

C_1 = {*A, A, B, C, C, C, D, D, E, E*} Set 1

C_2 = {*A, A, A, A, C, C, D, E, E, F, F*} Set 2

Therefore, the null hypothesis H_0 considers that the lists of strategies C_1 and C_2 are samples from the same population and, if it is not possible to reject H_0, it is accepted that the choice of business strategy is not a determiner of the level of competitiveness of a company. The aim of the presented work is to study how to answer this question through statistical tests.

This type of test may be done by determining whether the sets of values f_j and g_j can be considered as coming from the same population, where f_j and g_j are the distributions of the frequencies with which the strategies $j = 1, 2, ..., m$ appear in Group I and Group II of companies, respectively, so that $\sum_{j=1}^{m} f_j = n_1$ and $\sum_{j=1}^{m} g_j = n_2$. For the case of Chart 1, f_j and g_j assume the values expressed in Table 1.

3 Non-parametric tests and the problem of similarity of strategies

Non-parametric statistics include a large number of inference techniques whose preponderant factors are the few assumptions regarding how the data were generated. Normally, they only require the samples to be independents or the data to be obtained at random.

The fundamental problem in non-parametric statistics is determining, from the data of a sample, the probability value ρ (tail value) that will lead to the decision whether to accept the null hypothesis, which can be done in two ways:

a) Using the equation $\rho = P(X \geq x_{cal})$, where X represents a known probability distribution and x_{cal} is a value calculated from a (statistical) function of the sample data, so that $x_{cal} \in X$; or

b) Using the equation $\rho = \sum_{i=1}^{r} p_i$, where p_i, for i=1, is the probability of that configuration of values occurring reflected by the sample and p_i, $i = 2, ..., r$ is the probability of any one of the other possible $(r - 1)$ occurring, more extreme than the original sample.

Small ρ values (normally lower than $\alpha = 0.05$) indicate that the null hypothesis (H_0) should be rejected. Thus, it is vitally important to determine the value of ρ as accurately as possible.

The way that the ρ value is calculated divides non-parametric tests into two classes: *approximate tests* (or asymptotic tests), when ρ is determined as in (a), described above, and *exact tests*, when ρ is calculated as in (b). When the first way is chosen, for the obtained ρ value to be reliable, it is necessary to be certain that the test variable x_{cal} reproduces, with good approximation, a distribution element of X. A requisite condition for this is that the size of the sample should be sufficiently large. For this reason, they are called asymptotic tests. On the other hand, using method (b) there is the exact value for each p_i, which accounts for the origin of the term *exact test*.

A very common problem in statistical inference is determining, for a given level of test α, i.e., with the certainty of $(1 - \alpha)$, whether differences observed in two samples mean that the corresponding populations really differ from one another, which would lead to the rejection of the null hypothesis H_0, coinciding with the problem of interest to the *FWC* model.

The first tests developed in non-parametric statistics belong to the class of asymptotic tests. Lehmann (1975) attributes to John Arbuthnot (1710) the first work in the field through the presentation of the sign test, the purpose of which was to verify whether two samples stem from the same population, applying it to problems with ordinal variables. For a discussion on the types of variables (ordinal or categorized), several works can be consulted, including that of Siegel & Castellan (2006)

Pearson (1900) made a significant advance towards the creation of non-parametric tests applied to nominal or categorized variables. He demonstrated that the statistical test based on the sum of *m* samples formed by the differences between the observed frequency and expected frequency of variables distributed into *m* categories, when generated from a multinomial, hypergeometric or Poisson distribution, have a chi-square distribution providing the sample size is sufficiently large. This resulted in one of the most important asymptotic non-parametric tests (chi-square), applicable to a wide range of problems with categorized variables.

In the mid-twentieth century, non-parametric methods applied to problems with ordinal variables were given a boost by an article by Wilcoxon (1945), which presented a test based on the sum of ranks of two samples to verify whether they were extracted from the same population. Later, Mann & Whitney (1947) developed a more adequate procedure, which resulted in the Wilcoxon-Mann-Whitney test (Mann, Whitney and Wilcoxon, and others, independently proposed non-parametric tests that are essentially identical)

Other important works on non-parametric statistics that also addressed ordinal variables are those of Friedman (1937), Pitman (1937a, b, c), Kendall (1938), Smirnov (1939), Wald & Wolfowitz (1940), Kruskal & Wallis (1952), and Chernoff & Savage (1958).

From these works the following non-parametric tests were derived, and could be applied to the problem in question: sign test; Wilcoxon's sign rank test (1945);

Wilcoxon-Mann-Whitney rank sum test; chi-square, median test and the *t*-test for paired dataset. However, these tests are inadequate for addressing problems with small samples and categorized variables, as shown in their application to the data in Table 2.

Intuitively, it is difficult not to accept that there is no distinction between the two samples, as in six of the eleven classes there is a considerable difference between variables f_j and g_j.

In the sign test, as Respondent *A* surpasses *B* in six of the eleven requirements and is surpassed in three (there is a draw), a tail value equal to 0.254 is obtained, proving to be the true H_0. The Wilcoxon test provides a tail value of $\rho = 0.062$, for $T^+ = 51$ and n=11 and, through the Wilcoxon-Mann-Whitney test, for the variable of the test z=1.04 is obtained, which provides a two-tailed value equal to 0.298. When the median test is applied, for the respective contingency table, a chi-square value is obtained equal to $\chi^2_{cal} = 1.692$, showing that there is no distinction between the respondents (tail value $\rho = P[\chi^2_1 > 1.692]$=0.193). If we apply the *t*-test for paired dataset a two-tailed value of $\rho = 0.061$ is obtained. Finally, if we apply the chi-square test, we will obtain a tail value of $\rho = 0.675$.

Therefore, all the tests led to the conclusion that they would be the opposite of what was expected. This happened because for a statistical test to function adequately for the problem in question, the respective test variable X_{cal}, calculated from the data of the two samples, to be used to determine $\rho = P[X \geq X_{cal}]$, must have three properties. They are: a) considering the extent of the difference observed in each pair of values related to each class of random variable; b) accumulating the differences in opposite senses observed in different classes (stop one from annulling the other); and c) being adjusted to a known probability distribution *X*.

The only test among those applied with the first two properties is the chi-square. However, to meet the third requirement, it is necessary for at least 80% of the cells to have a frequency greater than 5 and no cell with a frequency less than 1 (Siegel & Castellan, 2006), which does not occur in the data in Table 3.

The chi-square can often fail if the values contained in the cells are sparse or have strong imbalance (see example in Section 7).

As an alternative to the chi-square, when the previous conditions are not met, the exact tests emerge. Fisher's test, proposed in 1925 (Fisher, 1970), was the first of these and is applicable to two samples of variables with two categories (tables with $l = 2$ lines and $c = 2$ columns). This test was later extended to tables with $l > 2$ and $c > 2$ by Freeman & Halton (1951). However, its application requires great computational effort, principally if the number of classes is large (Sprent & Smeeton, 2000, p. 322). In these cases, appropriate software is required, such as the StatXact (2008).

The uncertainty of using the chi-square in problems with small samples and the difficulty involved in applying exact tests led the authors to propose a new non-parametric test to address problems with small samples of categorized variables and conduct comparative studies on the performance of these three tests, i.e., the capacity to decide the null hypothesis H_0 correctly.

In the following section, the theory of the exact tests, especially that of Fisher, is presented, along with the procedure adopted by the *StatXact* software for this class of problem, with the main object being to show the difficulties involved in solving problems with small samples whose variables assume more than two categories.

4 Exact tests based on permutation theory

To exemplify the application of Fisher's exact test to tables with a 2x2 dimension, consider Tables 3a-c, in which Group I refers to the male sex and Group II to the female sex.

In the upper line of each of these tables are the frequencies of people with a height of 1.80m or taller. In the lower line, there are the frequencies of people who are under 1.80m tall. These data were obtained from a sample of eight men and nine women. The idea is to gauge, based on this small sample, whether men are taller than women. Consider that the hypothesis H_0 establishes equality of height and the alternative hypothesis H_1 establishes that men are taller than women. To apply Fisher's exact test to this problem, the value of $\rho = \sum_{i=1}^{r} p_i$ is determined, where p_i is the probability of an equal or more extreme situation occurring (in the sense of Hypothesis H_1) than that of Table 3a, maintaining the total fixed marginal values. Observe that the sample included six men who were taller than 1.80m and two who were shorter. As the

Table 2. Data for the application of the tests available in the literature.

Sample	*j*	1	2	3	4	5	6	7	8	9	10	11
A_1	f_j	5	4	5	4	5	4	4	4	4	4	5
A_2	g_j	2	1	2	1	2	1	5	5	4	5	5

Table 3. Data to exemplify Fisher's exact test.

Groups			Groups			Groups		
I	II		I	II		I	II	
6	3	9	7	2	9	8	1	9
2	6	8	1	7	8	0	8	8
8	9	17	8	9	17	8	9	17
(a)			(b)			(c)		

Source: Prepared by the authors.

test is unilateral, (due to the alternative hypothesis H_1), there are two other more extreme situations than that of Table 3a with fixed marginal values, which are represented in Tables 3b and 3c.

The exact probability of observing a particular set of frequencies in a 2x2 table, when the marginal totals are considered fixed, is given by the hypergeometric distribution, resulting in ρ = 0.109, obtained from the sum of $p_{(a)}$, $p_{(b)}$ and $p_{(c)}$, given by Equations 2, 3 and 4, respectively:

$$p_{(a)} = \frac{9!\ 8!\ 8!\ 9!}{17!\ 6!\ 3!\ 2!\ 6!} = 0.0968 \quad (2)$$

$$p_{(b)} = \frac{9!\ 8!\ 8!\ 9!}{17!\ 7!\ 2!\ 1!\ 7!} = 0.0012 \quad (3)$$

$$p_{(c)} = \frac{9!\ 8!\ 8!\ 9!}{17!\ 8!\ 1!\ 0!\ 8!} = 0.0004 \quad (4)$$

In this case, as ρ > 0.05, it is not possible to reject H_0 with a level of certainty of 95%.

an example will now be presented to illustrate how the exact test is applied to tables with $l > 2$ and $c > 2$.

Consider the data in Table 4 as representing the number of executives that belong to four business units of a large corporation who have been given high, average and low evaluations in an executive promotion program. Based on this small sample, is it possible to conclude that Business Unit *A* has the most capable executives (alternative hypothesis H_1)?

If the chi-square tests were applied, the constructed statistic would have $(l-1)\times(c-1) = 6$ degrees of freedom and would supply χ^2=11.555. As $P(\chi^2_6 > 11.555)$=0.0726, it would not be possible to reject the null hypothesis H_0 with 95% certainty and affirm that Business Unit *A* has more capable executives.

To apply the exact test, all the possible tables are generated from the configuration of the sample data, maintaining fixed marginal values. The tables that originate values of $\chi^2 \geq 11.555$ represent more extreme situations that of the original sample and thus contribute with their respective values of *p* to compose the value of ρ. For instance, Tables 5a and 5b are two possible arrangements obtained from Table 4. The first is χ^2=14.676, and should be considered a more extreme situation that that of the original model. Thus, its respective value of *p* contributes to the determination of ρ. Meanwhile, Table 5b provides χ^2=9.778, and its corresponding value of *p* does not contribute to the calculation of ρ.

Table 4. Result of the evaluation of executives.

Evaluation Level	Business Units				Total
	A	B	C	D	
High	5	2	2	0	9
Average	0	1	0	1	2
Low	0	2	3	4	9
Totals	5	5	5	5	20

The generalization of the calculation of probability *p* of a particular set of frequencies for a table with *l* lines and *c* columns, by Freeman & Halton (1951), is made using Equation 5, where $n_{i,o}$ is the marginal value of the line *i*, $n_{o,j}$ is the marginal value of the column *j*, n_{ij} is the value contained in the cell (*i, j*) and *n* is the sum of the values of all the cells:

$$p = \frac{\prod_i (n_{i,o})! \prod_j (n_{o,j})!}{n! \prod_{i,j} (n_{ij})!} \quad (5)$$

In the application of the exact test to tables of dimension $l \times c$, all the possible tables from the originating data of the sample must be represented. It is the representation of these tables that generally requires considerable computational effort.

This type of problem can be solved by software such as the StatXact (2008). For this particular case, this software arrives at ρ = 0.0398, which, contradicting the result of the chi-square test, leads to the rejection of the null hypothesis H_0 with 95% certainty.

5 Test based on the difference between two uniform distributions

In this section, a new non-parametric test is presented for the problem in question. The test statistic is given by the difference between two uniform distributions of probabilities.

Table 5. Two permutations of the results of the evaluation of the executives.

5	**2**	**2**	**0**	**9**		**4**	**3**	**2**	**0**	**9**
0	0	0	2	2		1	0	0	1	2
0	3	3	3	9		0	2	3	4	9
5	5	5	5	20		5	5	5	5	20
		(a)						(b)		

Let $j = 1, 2, ..., k, k \geq m$ be the index of the alternatives that a categorized random variable C can assume, and let $P = \{p_j, j = 1, 2, ..., k\}$ and $Q = \{q_j, j = 1, 2, ..., k\}$ be the true distributions of probabilities of this variable in two different populations P_1 and P_2 (e.g., more competitive and less competitive companies). Consider the functions expressed by Equations 6 and 7:

$$p_j' = p_j / [(p_j + q_j)/2] / \sum_{j=1}^{k} \{p_j / [(p_j + q_j)/2]\} \quad (6)$$

$$q_j' = q_j / [(p_j + q_j)/2] / \sum_{j=1}^{k} \{q_j / [(p_j + q_j)/2]\} \quad (7)$$

Then, if $p_j = q_j$, for all $j = 1, 2, ..., k$, it can easily be seen that $p_j' = q_j' = 1/k$, for all j, i.e., if P and Q have the same distribution of probabilities, then functions p_j and q_j convert the distribution of strategies j for both populations of companies into a uniform distribution with probability equal to $1/k$ for all j. This shows that the proposed test, which is in essence based on determining the difference for $|p_j' - q_j'|$, is convergent.

Now let f_j and $g_j, j = 1, 2, ..., m,$ be the frequencies that the random variable C assumes in two samples A_1 and A_2 of sizes n_1 and n_2 extracted from populations P_1 and P_2, respectively. As f_j / n_1 and g_j / n_2 are fair estimates for p_j and q_j, respectively, if A_1 and A_2 are samples from the same population, then the Equations 8 and 9 must have values close to $1/m$, for all $j = 1, 2, ..., m$, for any values of n_1 and n_2. This fact motivated the proposition of this test for the case of small samples, despite being an asymptotic test.

$$r_j = \{(f_j / n_1) / [(f_j + g_j)/(n_1 + n_2)]\} / \sum_{j=1}^{m} \{(f_j / n_1) / [(f_j + g_j)/(n_1 + n_2)]\} \quad (8)$$

$$s_j = \{(g_j / n_2) / [(f_j + g_j)/(n_1 + n_2)]\} / \sum_{j=1}^{m} \{(g_j / n_2) / [(f_j + g_j)/(n_1 + n_2)]\} \quad (9)$$

Now, consider the statistic $D = \sum_{j=1}^{m} |u_j - v_j|$, where u_j and v_j are relative frequencies of the variable $j = 1, 2, ..., m,$ so that $\Pr(j) = 1/m$, for all j. This variable is only a little sensitive to the variation of the number of elements in the sample (at least for small variations, which always occurs when dealing with small samples, as is the case in question). However, it depends on the value of m, as it stems from the sum of m samples, each one given by the difference between two uniform variables. The distribution of probabilities of this statistic is not known. Nevertheless, it is possible, through simulation, to construct its histogram for diverse values of m and, from each of these histograms, determine D_α, where D_α is the value of D that leaves α% of the data to its right.

With the aid of this information, it is possible to determine whether the lists of strategies A_1 and A_2 stem from the same population (Hypothesis H_0). It is sufficient to calculate the statistic $D_{cal} = \sum_{j=1}^{m} |r_j - s_j|$ from the values of f_j and g_j originating from A_1 and A_2, respectively, and compare with the value of D_α. If $D_{cal} > D_\alpha$, Hypothesis H_0 can be rejected with a level of certainty (1-α).

Observe that the variable D_{cal} (like D) is defined in the interval [0, 2]. When $f_j = g_j$, for all $j = 1, 2, ..., m$, then $D_{cal} = 0$, which provides the maximum certainty that both sets A_1 and A_2 originate from the same population. Now, when, for each $j = 1, 2, ..., m$, ($f_j = 0$, $g_j > 0$) or ($f_j > 0$, $g_j = 0$), which means that each group of companies declared different sets of strategies and therefore the intersection of sets A_1 and A_2 is empty, then $D_{cal} = 2$, which provides the maximum certainty of rejection for the null hypothesis H_0.

5.1 Determining the value of D_α

The value of D_α was determined from the histogram of the variable D, constructed through a computer simulation process. This procedure is illustrated below for the case of $m = 6$, $n_1 = n_2 = 12$.

Step 1. Establish the following correlation according Table 6, where *RN* is a uniform random number in the interval [0, 1].

Step 2. Generate n_1 uniform random numbers (*RN*) in the interval [0, 1] for the first sample and other n_2 numbers for the second sample, and obtain sets A_1 and A_2, i.e., values of f_j and g_j. For $n_1 = n_2 = 12$, a possible result is shown in columns f_j and g_j of Table 7, where, among the twelve values randomly selected for sample A_1, two fell in the interval [0, 1/6), and for sample A_2, three values fell in the same interval, thus originating $f_1 = 2$ and $g_1 = 3$.

Step 3. Determine, for each generated sample A_1 and A_2, $D = \sum_{j=1}^{m} |u_j - v_j|$, where $u_j = (f_j / n_1)$, $v_j = (g_j / n_2)$, as shown in Table 7, which arrives at $D = 0.333$ for this example.

Step 4. Repeat Steps 1 to 3 10000 times, generating 10000 ordered values for D, and identify the value of D_α, for significance levels $\alpha = 0.01$ and $\alpha = 0.05$ ($D_{0.05}$ is given by the value of D, which leaves 500 values to its right, and $D_{0.01}$ is given by the value of D that leaves 100 values to its right). Table 8 shows

Table 6. Correlation between uniform random number and the classes of variables.

***RN* in the interval**	**Variable *C***
[0, 1/6)	A
[1/6, 2/6)	B
[2/6, 3/6)	C
[3/6, 4/6)	D
[4/6, 5/6)	E
[5/6, 1]	F

Table 7. Application of the test of the difference of two uniform distributions.

Strategy (*j*)	f_j	g_j	u_j	v_j	$\|u_l - v_j\|$
A	2	3	0.167	0.250	0.083
B	1	0	0.083	0.000	0.083
C	3	3	0.250	0.250	0.000
D	2	1	0.167	0.083	0.083
E	2	2	0.167	0.167	0.000
F	2	3	0.167	0.250	0.083
Sum	12	12	1.000	1.000	0.333

Table 8. Critical values of D_{α}

α	***m***					
	3	**4**	**5**	**6**	**7**	**8**
0.05	1.143	1.250	1.200	1.167	1.143	1.125
0.01	1.429	1.500	1.400	1.333	1.286	1.250

the critical values of D_{α}, for different values of m and $\alpha = 0.05$ and $\alpha = 0.01$.

Applying the test to the data in Table 2, $D_{cal} = 0.493$ is obtained. As in this example $m = 6$, it can be concluded that the null hypothesis H_0 cannot be rejected and it must be accepted that the two groups of companies adopt similar sets of strategy.

6 Study of the power of the tests

The effectiveness of the exact tests, the chi-square and the proposed test, was evaluated by analyzing the power curve, supplying the probability of the acceptance (*Pa*) of the null hypothesis (H_0) due to the level of similarity of the two samples.

The power curve was raised using computer simulation for the level of similarity between the samples, defined by the parameter referred to as the *degree of symmetry* (*DS*) of the distributions of samples A_1 and A_2, varying in the interval [0, 1] and given by Equation 10, where p_j and q_j are the probabilities of the categorized variable originating from samples A_1 and A_2 for all $j = \{1, 2, ..., m\}$.

$$GS = (\sum_{j=1}^{m} |p_j - q_j|) / 2 \quad (10)$$

Defining appropriate values for p_j and q_j, through simulations, samples from populations with the following degrees of symmetry were obtained $DS = \{0.0; 0.2; 0.4; 0.6$ and $0.8\}$. Observe that if $p_j = q_j$ for all j, Equation 2 provides $DS = 0$, and the samples obtained by simulation for this case will be from the same population. On the other hand, if $p_j = 0$ when $q_j \neq 0$, for all j, then $DS = 1$, creating configurations with samples from totally different populations.

Computer tests were conducted for the following six configurations of problems identified by the sets of values of (m, n_1, n_2): (3, 7, 7), (4, 8, 8), (5, 10, 10), (6, 12, 12), (7, 14, 14) and (8, 16, 16). For each of these six cases and for each of the five values of *DS* mentioned above, the probability of acceptance *Pa* was determined according to the exact test, the chi-square and the proposed test.

For this purpose, 100 problems were generated for each of the six sets of values (m, n_1, n_2) and the five degrees of symmetry (*DS)*. The value of *Pa* for a determined test and for a given set of values (m, n_1, n_2) and a given value of *DS* could then be identified by directly counting the number of problems in which there would be acceptance of the H_0.

For all the tests, a significance level $\alpha = 0.05$ was adopted. Thus, the acceptance of H_0 occurred whenever $\rho = P[X > X_{cal}] > \alpha$, where X is the test variable and X_{cal} is the value of the statistic of the test, or whenever $X_{cal} < X_{crit}$, where X_{crit} is such that $P[X > X_{crit}] = \alpha$, which is the same thing viewed in two ways.

In all, 3000 problems were tested, 100 for each combination $[(m, n_1, n_2); GS]$, and each was solved using the three tests.

The configuration of each problem, i.e., values of f_j and g_j for both samples was obtained as described in Steps 1 and 2 of the procedure to determine D_{α}, presented in Section 5.

From this curve, raised by computer simulation, the following indicators could be extracted for a comparative analysis of the tests:

a) Risk α, which is the probability of committing a Type I error (rejecting the null hypothesis when it is true), given by $\alpha = (1\text{-}Pa)$, for $DS = 0$;

b) Average of risks β given by the average of *Pa* for the four values of $DS > 0$, where β is the probability of committing a Type II error (accepting a false null hypothesis); and

c) Characteristic indicator of the power curve (*CI*), determined by the relationship $(Slope)_{0.50}/(DS)_{0.50}$, where $(Slope)_{0.50}$ is the slope of the curve at point $(DS)_{0.50}$, with $(DS)_{0.50}$ being the value of

DS that originates a probability of acceptance of 50%.

The value of $(Slope)_{0.50}$ was determined by Equation 11.

$$(Slope)_{0.50} = -\frac{(DS)_{0.6} - (DS)_{0.4}}{100.(0.6 - 0.4)} \quad (11)$$

As it is a downward curve, the negative sign is introduced to make the result of the slope positive. The denominator was multiplied by 100 to represent it on a more adequate scale (interval [1 to 10]). The values of $(DS)_{0.40}$ and $(DS)_{0.60}$ were obtained by visual inspection of the graph of the power curve generated by the five points (*GS*, *Pa*).

The two parameters $(Slope)_{0.50}$ and $(DS)_{0.50}$ are frequently used to evaluate the discriminant power of quality inspection plans. The higher the value of $(Slope)_{0.50}$ and the lower the value of $(DS)_{0.50}$, the greater the power of the plan, or the power of the statistical test, in the present study. Thus, the *CI* expresses in a single indicator the properties of both (the higher their value, the greater the power of the test) and can dispel doubts that may remain from the application of the α and β risk indicators.

Studies on the performance of statistical tests adopts only the risk indicators, a point in question being the case of Tanizaki (1997). Thus, the use of a new indicator (*CI*) with the property outlined above makes a contribution to this type of study.

Tables 9a to 9f show the results obtained from the trials with the exact tests (solution obtained by the *StatXact*), chi-square (*Chi-Squ*) and presented test (Uniform). The values of *Pa* are expressed in percentage form, as they correspond directly to the number of problems in which there was acceptance of H_0, out of 100 problems tested for each value of *DS*. The meaning and form of obtaining the values of *CI*, α, and β Average, shown in in Tables 9a-f will be explained in the following section.

Table 9a. Results for $m = 3$, $n_1 = n_2 = 7$.

Teste	Probability of acceptance (*Pa*) - Percentage					*CI and Risks* (%)		
	***DS*=0**	***DS*=0.2**	***DS*=0.4**	***DS*=0.6**	***DS*=0.8**	**CI**	**α**	**β Average**
Exact	98	94	80	24	0	5.6	2	50
Chi-Squ	95	89	73	16	0	5.9	5	45
Uniform	97	92	77	20	0	6.6	3	47

Source: Prepared by the authors.

Table 9b. Results for $m = 4$, $n_1 = n_2 = 8$.

Teste	Probability of acceptance (*Pa*) - Percentage					*CI and Risks* (%)		
	***DS*=0**	***DS*=0.2**	***DS*=0.4**	***DS*=0.6**	***DS*=0.8**	**CI**	**α**	**β Average**
Exact	97	94	78	51	12	2.3	3	59
Chi-Squ	96	93	78	50	11	2.3	4	58
Uniform	96	93	82	58	17	1.3	4	63

Source: Prepared by the authors.

Table 9c. Results for $m = 5$, $n_1 = n_2 = 10$.

Teste	Probability of acceptance (*Pa*) - Percentage					*CI and Risks* (%)		
	***DS*=0**	***DS*=0.2**	***DS*=0.4**	***DS*=0.6**	***DS*=0.8**	**CI**	**α**	**β Average**
Exact	96	92	78	34	3	4.2	4	52
Chi-Squ	97	94	83	38	5	4.1	3	55
Uniform	95	92	84	39	4	3.4	5	55

Source: Prepared by the authors.

Table 9d. Results for $m = 6$, $n_1 = n_2 = 12$.

Teste	Probability of acceptance (*Pa*) - Percentage					*CI and Risks* (%)		
	***DS*=0**	***DS*=0.2**	***DS*=0.4**	***DS*=0.6**	***DS*=0.8**	**CI**	**α**	**β Average**
Exact	90	94	71	28	6	4.3	10	50
Chi-Squ	92	96	77	38	7	3.5	8	55
Uniform	91	85	74	39	9	3.9	9	52

Source: Prepared by the authors.

Table 9e. Results for $m = 7$, $n_1 = n_2 = 14$.

Teste	Probability of acceptance (*Pa*) - Percentage					*CI and Risks* (%)		
	DS=0	*DS*=0.2	*DS*=0.4	*DS*=0.6	*DS*=0.8	CI	α	β Average
Exact	90	94	67	32	3	3.5	10	49
Chi-Squ	95	95	73	42	3	2.7	5	53
Uniform	94	96	70	46	3	2.3	6	54

Source: Prepared by the authors.

Table 9f. Results for $m = 8$, $n_1 = n_2 = 16$.

Teste	Probability of acceptance (*Pa*) - Percentage					*CI and Risks* (%)		
	DS=0	*DS*=0.2	*DS*=0.4	*DS*=0.6	*DS*=0.8	CI	α	β Average
Exact	95	88	68	21	1	4.9	5	45
Chi-Squ	97	94	79	26	2	5.8	3	50
Uniform	91	91	72	32	5	3.9	9	50

Source: Prepared by the authors.

7 Analysis of the results and conclusions

The effectiveness of the tests was evaluated by risks α and β and the characteristic indicator of the power curve (*CI*).

Risk α for each configuration of problem (m, n_1, n_2) is given in percentage form in the respective Table 9 by the value (100- *Pa*) for the column *DS*=0, as the value of *Pa* corresponds, among the 100 trials conducted, to the number in which the test led to the right decision, i.e., accepting the H_0 when it is true. Meanwhile, risk β, also in percentages, is given by the average of the values of *Pa* for all *DS*={0.2, 0.4 0.6, 0.8}, i.e., the probability of accepting H_0 when it is not true (sample with a degree of symmetry other than zero).

The values of $(Slope)_{0.50}$, for each configuration (m, n_1, n_2), were calculated using Equation 3. These three analysis parameters are shown in Table 10.

Analyzing risks α and β in Table 10, the chi-square tests has the lowest risk α of the three and risk β between the other two. However, regarding the *CI* indicator, it had the worst performance of the three.

The proposed test shows risks α and β, and the characteristic indicator (*IC*), similar to those of the exact test. This shows that both have a very similar performance.

Table 11 shows the number of problems that each test decided on correctly for the 3000 trials that took place. The exact test made the most right decisions (1753 times) while the proposed test had a slightly inferior performance to the other two.

This analysis allows us to conclude that the exact and proposed tests had very similar performances, and that the chi-square surpasses both, at least as an instrument for decision, when the null hypothesis is true. To a certain extent, this is an unexpected conclusion, when it comes to problems with small samples. In view of this, would the chi-square a valid alternative to the exact test?

Table 10. Summary of the parameters of evaluation of the effectiveness of the tests.

Parameter	*m*	Test		
		Exact	Chi-square	Uniform
Risk α (%)	3	2.0	5.0	3.0
	4	3.0	4.0	4.0
	5	4.0	3.0	5.0
	6	10.0	8.0	9.0
	7	10.0	5.0	6.0
	8	5.0	3.0	9.0
	Average value	5.7	4.7	6.0
Risk β average (%)	3	49.5	44.5	47.3
	4	58.8	58.0	62.5
	5	51.8	55.0	54.8
	6	49.8	54.5	51.8
	7	49.0	53.3	53.8
	8	44.5	50.3	50.0
	Average value	50.5	52.6	53.3
CI	3	5.9	6.6	5.6
	4	2.3	1.3	2.3
	5	4.1	3.4	4.2
	6	3.5	2.9	4.3
	7	2.7	2.3	3.5
	8	5.8	3.9	4.9
	Average value	3.9	3.1	4.1

Source: Prepared by the authors.

Considering the example of the data in Table 12, this is not always the case. Applying the exact test to the data in this table, by the *StatXact*, ρ = 0.0013 is obtained, showing that the three samples do not belong to the same population. In turn, the chi-square gives a value of ρ= 0.1342, clearly showing that for small

Table 11. Number of problems with the right decision.

DS	Testes		
	Exact	**Chi-Squ**	**Uniform**
0.00	566	572	564
0.20	44	39	51
0.40	158	137	141
0.60	410	390	366
0.80	575	572	562
All	1753	1710	1684

Source: Prepared by the authors.

Table 12. Example of a problem with three samples.

Sample	Values								
A	0	7	0	0	0	0	0	1	1
B	1	1	1	1	1	1	1	0	0
C	0	8	0	0	0	0	0	0	0

Source: StatXact (2003).

samples with a strong imbalance, as in this example, this test does not work well. Table 5 provides another example of this phenomenon. Thus, its generalized use leads to unreliable decisions, explaining the need to seek alternative tests.

How does the proposed test behave with this type of sample?

To answer this question, it is initially necessary to observe that although the proposed test is intended for problems with two samples, it is also possible to solve problems with more samples. All that is required is to apply it to the different combinations of samples two by two. Applying the uniform test to the data in Table 12 considering two samples at a time (observe that it is necessary to eliminate the columns that contain zeros in both samples), values are obtained for D_{cal} equal to 1.959, 1.622 and 1.964 for the combinations *A*/*B*, *A*/*C* and *B*/*C* of samples, respectively. As the maximum value of *Dcal* is 2.0, the test indicates with a high degree of certainty that sample *B* is from a different population from the others, which the chi-square failed to identify.

If we now apply the proposed test to the data in Table 5, the values obtained for D_{cal} are equal to 1.750, 1.556 and 2.000 for samples *A*/*B*, *A*/*C* and *A*/*D*, respectively. As $D_{\alpha=0.01}$ =1.429, for *m*=3 (case of Table 4), it can be concluded, with a high degree of certainty, that Business Unit *A* has the most capable executives.

These two examples show that the best alternative to the exact test, which is very difficult to apply, is the proposed test rather than the chi-square. The latter, despite having shown a good performance in the set of tests, can fail in accordance with the instance of the problem.

References

Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27(325-336), 186-190. http://dx.doi.org/10.1098/rstl.1710.0011.

Chernoff, H., & Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric tests. *Annals of Mathematical Statistics*, 29(4), 972-994. http://dx.doi.org/10.1214/aoms/1177706436.

Contador, J. C. (2008). *Campos e armas da competição: novo modelo de estratégia*. São Paulo: Sant Paul.

Fisher, R. A. (1970). *Statistical methods for research workers*. 14. ed. Edinburgh: Oliver and Boyd.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency goodness-of-fit and other problems of significance. *Biometrika*, 38(1-2), 141-149. http://dx.doi.org/10.1093/biomet/38.1-2.141. PMid:14848119.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675-701. http://dx.doi.org/10.1080/01621459.1937.10503522.

Hamel, G., & Prahalad, C. K. (1995). *Competindo pelo futuro*. Rio de Janeiro: Campus.

Kendall, M. G. (1938). A new measure correlation. *Biometrika*, 30(1-2), 81-93. http://dx.doi.org/10.1093/biomet/30.1-2.81.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583-621. http://dx.doi.org/10.1080/01621459.1952.10483441.

Lehmann, E. L. (1975). *Nonparametrics: statistical methods based on ranks*. San Francisco: Holden-Day.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1), 50-60. http://dx.doi.org/10.1214/aoms/1177730491.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302), 157-175. http://dx.doi.org/10.1080/14786440009463897.

Pitman, E. J. G. (1937a). Significance tests which may be applied to sample from any populations. *Journal of the Royal Society*, 4, 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to sample from any populations - II. The correlation coefficient test. *Journal of the Royal Society*, 4, 225-232.

Pitman, E. J. G. (1937c). Significance tests which may be applied to sample from any populations - III. The analysis of variance test. *Biometrika*, 29, 322-335.

Siegel, S., & Castellan, N. J., Jr. (2006). *Estatística não-paramétrica para ciências do comportamento*. 2. ed. Porto Alegre: Artmed.

Smirnov, N. V. (1939). Estimate of difference between empirical distribution curves in two independent samples. *Moscow University Mathematics Bulletin*, 2(2), 3-4.

Sprent, P., & Smeeton, N. C. (2000). *Applied nonparametric statistical methods*. 3. ed. New York: Chapman & Hall.

StatXact. (2003). *Software for small-sample categorical and nonparametric data: user manual, Versão 6*. Cambridge.

StatXact. (2008). *Software for small-sample categorical and nonparametric data*. Cambridge. Recuperado em 01 de dezembro de 2008, de http://www.cytel.com/products/statxact/

Tanizaki, H. (1997). Power comparison of non-parametric tests: small sample properties from Monte Carlo experiments. *Journal of Applied Statistics*, 24(5), 603-632. http://dx.doi.org/10.1080/02664769723576.

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 11(2), 147-162. http://dx.doi.org/10.1214/aoms/1177731909.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83. http://dx.doi.org/10.2307/3001968.