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

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

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

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

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.

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.*

1 Introdução

A motivação para este trabalho surgiu da necessidade de criar um teste estatístico de fácil aplicação para auxiliar as pesquisas que embasaram o desenvolvimento do modelo de Campos e Armas da Competição – CAC (Contador, 2008), cujo interesse era (entre outras coisas) verificar se a estratégia de negócio adotada pela empresa é um fator determinante para sua competitividade. Em suas pesquisas, o autor desse modelo colhia uma pequena amostra de empresas as quais eram divididas em dois grupos, um reunindo as mais competitivas e outro, as menos competitivas, e o teste presta-se para verificar se ambos os grupos adotam estratégias de negócio semelhantes (hipótese nula H_0).

Qualquer problema que apresente as seguintes características pode utilizar o teste aqui proposto:

a) presença de dois grupos distintos, *I* e *II* (por exemplo, empresas mais competitivas e empresas menos competitivas), representando amostras de populações maiores, com n_1 e n_2 elementos em cada grupo, onde n_1 e n_2 são valores pequenos;

[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, Brasil, 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, Brasil, e-mail: elfsenne@feg.unesp.br

Recebido em Out. 2, 2014 - Aceito em Dez. 18, 2015
Suporte financeiro: O primeiro autor agradece o suporte financeiro do CNPq (DT 307363/2015-5). O segundo autor agradece o suporte financeiro do CNPq (grant 303339/2013-6).

b) a variável aleatória assume, para cada grupo ou amostra, valores de frequências em cada uma das m classes, $m > 2$, (vide Tabela 1), ou seja, a mensuração da variável aleatória é feita numa escala nominal ou categorizada com mais de duas categorias;

c) O número de classes ou de categoria que a variável aleatória pode assumir (valor de m) é moderado em relação aos valores de n_1 e n_2.

Observe-se que se a variável aleatória pudesse ser classificada em apenas duas categorias (duas estratégias, por exemplo) o problema poderia ser facilmente resolvido pelo teste exato de Fisher (vide seção 4), qualquer que fosse o tamanho n_1 e n_2 das amostras dos dois grupos.

Se, por outro lado, existissem mais de duas categorias para a variável aleatória, mas para cada classe um número suficientemente grande de indivíduos (o que geraria um problema com grandes amostras), a verificação da semelhança entre os dois conjuntos de respostas poderia ser feita também facilmente por meio do teste qui-quadrado, que pode falhar quando se tratar de pequenas amostras.

Os demais testes não paramétricos disponíveis (teste do sinal, teste de postos com sinal, teste da soma dos postos, teste da mediana e teste t para amostras pareadas) são inadequados, como será mostrado por meio de exemplos. Assim, para o caso de pequenas amostras e mais de duas classes para a variável aleatória, o problema torna-se de difícil solução.

Portanto, a única alternativa segura para tratar esse tipo de problema são os testes exatos, como, por exemplo, aquele apresentado em StatXact (2008), cuja solução baseia-se numa extensão do teste exato de Fisher (1970) proposta por Freeman & Halton (1951). Contudo, a implementação desse teste requer algoritmos específicos e, em alguns casos, exige grande esforço computacional, o que justifica a busca de novos testes para esses tipos de problema.

Em vista disso, este artigo apresenta um estudo comparativo do desempenho (capacidade de decidir H_0 corretamente) dos testes exato, qui-quadrado e de um novo teste baseado na diferença de duas distribuições uniformes, aqui proposto. A comparação da eficácia desses testes é feita por meio de três indicadores (riscos α e β e o indicador característico – IC) extraídos da sua curva de poder, a qual será construída por meio de simulação.

Tabela 1. Frequências das estratégias (*CC*) para os grupos de empresa.

CC	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

Os estudos aqui desenvolvidos estão voltados à tentativa de solução do problema de estratégia relacionado com o modelo CAC, motivo pelo qual são fornecidos na seção seguinte alguns conceitos sobre esse modelo, essenciais para entender o problema em questão. O objetivo deste artigo não é discutir ou apresentar o modelo CAC. Caso o leitor esteja interessado em aprofundar seus conhecimentos sobre ele, poderá consultar a referência fornecida.

Inúmeros outros problemas relacionadas à biologia, medicina, ciências sociais e humanas, apresentam as características anteriormente descrita e poderiam ser abordados pelas técnicas estatísticas aqui tratadas. Alguns exemplos de problemas diretamente relacionados às engenharias sociais são:

- Verificar se dois tipos distintos de funcionários (operadores de máquina e funcionários de escritório, por exemplo), em empresas de pequeno porte (com poucos funcionários), motivam-se de forma semelhante frente aos diversos fatores motivacionais, para permitir desenvolver um único programa de incentivo (ou serem incluídos num único programa);
- Verificar, por meio de pequena amostra, se empresas de setores distintos (transformação e serviços, por exemplo) valorizam as mesmas características dos seus executivos para universalizar os programas de desenvolvimento humano;
- Verificar se os executivos (que são em pequeno número) das diversas unidades de negócio de uma corporação apresentam capacidade de gestão semelhante;
- Verificar se dois processos produtivos distintos, pela análise de poucas peças, geram produtos com nível de qualidade semelhante para as diversas características (dimensões, acabamento, etc.).

Como principal resultado do trabalho, verificou-se que o teste propostos apresenta eficácia parecida ao do teste exato e se comporta muito bem em situações onde o teste qui-quadrado mostra-se falho (amostras pequenas, dados esparsos com forte desbalanceamento), sendo, portanto, uma real alternativa ao teste exato, cuja aplicação muitas vezes obriga lançar mão de *softwares* especiais com acesso restrito.

Na seção 3 apresenta-se uma breve discussão sobre os testes não paramétricos e uma análise crítica sobre a aplicação desses testes na solução do problema em questão (de estratégia). Na seção 4 apresenta-se o método de solução adotado pelo StatXact para

problemas com variáveis categorizadas. Na seção 5 é apresentado o desenvolvimento do teste proposto, baseado na diferença entre duas distribuições uniformes. Na seção 6 são apresentados os estudos realizados para avaliar o desempenho dos três testes (o proposto, o teste exato e qui-quadrado) e, na seção 7, as conclusões do trabalho. Nessa última seção mostra-se também como o teste proposto pode ser estendido a problemas com mais de duas amostras independentes e são apresentados dois exemplos em que o teste proposto apresenta clara vantagem em relação ao qui-quadrado.

2 Modelo de campos e armas da competição

Segundo o modelo CAC, as empresas centram sua estratégia competitiva de negócio em um dos 14 campos da competição (agregados em cinco macrocampos), embora possam adotar mais um ou dois campos coadjuvantes. Os campos da competição, segundo o modelo CAC, são os seguintes:

- *Macrocampo da competição em preço*: (1) preço propriamente dito, (2) condições de pagamento e (3) prêmio e/ou promoção;
- *Macrocampo da competição em produto, bem ou serviço*: (4) projeto do produto, (5) qualidade do produto e (6) variedade de modelos;
- *Macrocampo da competição em assistência*: (7) assessoramento tecnológico antes da venda, (8) atendimento durante a venda e (9) assistência técnica após a venda;
- *Macrocampo da competição em prazo*: (10) prazo de cotação/negociação e (11) prazo de entrega do produto;
- *Macrocampo da competição em imagem*: (12) do produto e da marca, (13) empresa confiável e (14) responsabilidade social (cívica e preservacionista).

A tese do modelo CAC sustenta que não é a escolha da estratégia competitiva que determina a competitividade da empresa, mas sim o correto alinhamento da sua competência essencial (*core competence*, segundo Hamel & Prahalad, 1995) ao campo escolhido para competir, seja ele qual for. Evidentemente, sustenta o modelo, deve-se escolher, para cada par produto/mercado, um daqueles campos que atendem ao interesse do mercado.

Para melhor entendimento do problema em questão, considere os dados do Quadro 1, extraídos de uma das pesquisas realizadas por Contador (2008), que apresenta um conjunto de 21 empresas, as quais, pelo grau de competitividade (*GC*) apresentado, foram divididas em dois grupos: o das empresas mais competitivas e o das menos competitivas. Para determinar o grau de competitividade da empresa *i* (GC_i), o modelo CAC normalmente utiliza a variação ocorrida num determinado período de tempo do faturamento ou da receita líquida dessa empresa.

A classificação de uma empresa *i* no grupo das mais ou das menos competitivas é feita, no modelo CAC, por meio do índice de Nihans (*N*). Para um grupo de *n* empresas, o índice de Nihans é calculado por meio da fórmula seguinte, expressa pela Equação 1:

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

Assim, se $GC_i \geq N$, então a empresa é classificada no grupo das mais competitivas, caso contrário, é classificada no outro grupo.

A coluna *CC* de cada grupo de empresa do Quadro 1 apresenta os códigos dos principais campos

Quadro 1. Classificação das empresas nos grupos das mais e das menos competitivas.

Grupo I: Empresas mais competitivas				Grupo II: Empresas menos competitivas			
Cód	Principal campo da competição / Denominação	*CC*	GC_i	Cód	Principal campo da competição / Denominação	*CC*	GC_i
E10	Imagem do produto e marca	A	1,51	E05	Variedade de modelos	D	0,82
E13	Prazo de entrega do produto	B	1,43	E11	Assistência após a venda	C	0,80
E17	Assistência após a venda	C	1,39	E06	Imagem do produto e marca	A	0,79
E19	Assistência após a venda	C	1,32	E12	Imagem do produto e marca	A	0,79
E21	Variedade de modelos	D	1,25	E04	Imagem do produto e marca	A	0,69
E02	Imagem do produto e marca	A	1,19	E14	Assistência antes da venda	F	0,62
E08	Projeto do produto	E	1,16	E16	Projeto do produto	E	0,54
E03	Assistência após a venda	C	1,14	E07	Imagem do produto e marca	A	0,47
E13	Projeto do produto	E	1,11	E09	Assistência antes da venda	F	0,38
E01	Variedade de modelos	D	1,07	E20	Projeto do produto	E	0,30
				E18	Assistência após a venda	C	0,25

Fonte: Contador (2008).

da competição declarados pela respectiva empresa. Dessa forma, as estratégias dos dois grupos de empresas podem ser representadas pelas listas C_1 (Conjunto 1 – empresas mais competitivas) e C_2 (Conjunto 2 – empresas menos competitivas):

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

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

Portanto, se a hipótese nula H_0 considera que as listas de estratégias C_1 e C_2 são amostras provenientes de uma mesma população e, se não for possível rejeitar H_0, se aceita que a escolha da estratégia de negócio não é determinante para o nível de competitividade da empresa. O objetivo deste trabalho é estudar como responder a essa questão por meio de testes estatísticos.

Esse tipo de teste é feito verificando-se se os conjuntos de valores f_j e g_j podem ser considerados provenientes de uma mesma população, na qual f_j e g_j são as distribuições das frequências com que as estratégias j = 1, 2, ..., m aparecem no Grupo I e no Grupo II de empresas, respectivamente, tal que $\sum_{j=1}^{m} f_j = n_1$ e $\sum_{j=1}^{m} g_j = n_2$. Para o caso do Quadro 1, f_j e g_j assumem os valores expressos na Tabela 1.

3 Testes não paramétricos e o problema da semelhança entre estratégias

A estatística não paramétrica agrega um grande número de técnicas de inferência cujo fator preponderante são as poucas suposições sobre como os dados foram gerados. Normalmente, exigem apenas que as amostras sejam independentes ou que os dados sejam obtidos aleatoriamente.

O problema fundamental em estatística não paramétrica é a determinação, a partir dos dados de uma amostra, do valor de probabilidade ρ (valor de cauda) que levará à decisão sobre aceitar ou não a hipótese nula, o que pode ser feito de duas maneiras:

a) por meio da expressão $\rho = P(X \geq x_{cal})$, na qual X representa uma distribuição de probabilidade conhecida e x_{cal} é um valor calculado a partir de uma função (estatística) dos dados da amostra, tal que $x_{cal} \in X$; ou

b) por meio da expressão $\rho = \sum_{i=1}^{r} p_i$, na qual p_i, para i = 1, é a probabilidade de ocorrer aquela configuração de valores refletida pela amostra e p_i, i = 2, ..., r, é a probabilidade de ocorrer qualquer uma das outras (r – 1) possíveis configurações mais extremas do que a da amostra original.

Valores de ρ pequenos (normalmente menores do que α = 0,05) indicam que a hipótese nula (H_0) deve ser rejeitada. Assim, é de crucial importância determinar da forma mais acurada possível o valor de ρ.

A forma pela qual o valor de ρ é calculado divide os testes não paramétricos em duas classes: *testes aproximativos* (ou assintóticos), quando ρ é determinado da maneira a anteriormente descrita, e *testes exatos*, quando ρ é calculado da maneira b. Quando se opta pela primeira maneira, para que se tenha confiança no valor obtido para ρ, deve-se ter certeza de que a variável de teste x_{cal} reproduz, com boa aproximação, um elemento da distribuição de X. Uma condição indispensável para isso é que o tamanho da amostra seja suficientemente grande, por isso são chamados de testes assintóticos. Por outro lado, pela maneira b tem-se o valor exato para cada p_i, e, portanto, para ρ, o que justifica a origem do termo *teste exato*.

Um problema muito comum em inferência estatística é determinar, para um dado nível do teste α, ou seja, com certeza de (1 – α), se diferenças observadas em duas amostras significam que as populações correspondentes são realmente diferentes entre si, o que levaria à rejeição da hipótese nula H_0, e que coincide com o problema de interesse do modelo CAC.

Os primeiros testes desenvolvidos em estatística não paramétrica pertencem à classe dos testes assintóticos. Lehmann (1975) atribui a John Arbuthnot (1710) o primeiro trabalho na área, pela apresentação do *teste do sinal* cujo objetivo é verificar se duas amostras provêm de uma mesma população e aplica-se a problemas com variáveis ordinais. Para uma discussão sobre tipos de variáveis (ordinais ou categorizadas), vide, por exemplo, Siegel & Castellan (2006)

Pearson (1900) deu um grande passo para a criação de testes não paramétricos aplicados a variáveis nominais ou categorizadas, demonstrando que o teste estatístico baseado na soma das m parcelas formadas pelas diferenças entre a frequência observada e a frequência esperada de variáveis distribuídas em m categorias, quando geradas de uma distribuição multinomial, hipergeométrica ou de Poisson, possui distribuição qui-quadrado, desde que o tamanho da amostra seja suficientemente grande. Esse resultado gerou um dos mais importantes testes não paramétricos assintóticos (qui-quadrado), aplicável em uma extensa classe de problemas com variáveis categorizadas.

Em meados do Século XX, os métodos não paramétricos aplicados a problemas com variáveis ordinais receberam grande impulso a partir do artigo de Wilcoxon (1945), que apresenta um teste baseado na soma dos postos de duas amostras para verificar se são extraídas de uma mesma população. Mais tarde, Mann & Whitney (1947) desenvolveram um procedimento mais adequado, o que originou a prova conhecida por teste de Wilcoxon-Mann-Whitney (Mann, Whitiney e Wilcoxon, entre outros, propuseram, independentemente, testes não paramétricos os quais são essencialmente iguais)

Outros importantes trabalhos iniciais em estatística não paramétrica que também abordam variáveis ordinais são Friedman (1937), Pitman (1937a, b, c), Kendall

(1938), Smirnov (1939), Wald & Wolfowitz (1940), Kruskal & Wallis (1952) e Chernoff & Savage (1958).

Desses trabalhos originaram-se os seguintes testes não paramétricos disponíveis que, aparentemente, poderiam ser aplicados ao problema em questão: teste do sinal; teste de postos com sinal de Wilcoxon (1945); teste da soma dos postos de Wilcoxon-Mann-Whitney; qui-quadrado, teste da mediana; e teste *t* para amostras pareadas. Porém, esses testes são inadequados para tratar o problema com pequenas amostras e variáveis categorizadas, como mostra sua aplicação nos dado da Tabela 2.

Intuitivamente, é difícil não aceitar que não haja distinção entre as duas amostras, uma vez que em 6 das 11 classes ocorre forte diferença entre as variáveis f_j e g_j.

Pelo teste do sinal, como o respondente *A* supera o *B* em 6 dos 11 quesitos e é superado em 3 quesitos (ocorreu um empate), obtém-se o valor de caudal igual a 0,254, mostrando ser H_0 verdadeira. O teste de Wilcoxon fornece valor de cauda $\rho = 0{,}062$, para $T^+ = 51$ e $n = 11$ e, pelo teste de Wilcoxon-Mann-Whitney, obtém-se para a variável do teste $z = 1{,}04$, o que fornece valor bicaudal igual a 0,298. Ao aplicar o teste da mediana obtém-se, para a respectiva tabela de contingência, valor do qui-quadrado igual a $\chi^2_{cal} = 1{,}692$, evidenciando não haver distinção entre os respondentes (valor de cauda $\rho = P[\chi^2_1 > 1{,}692] = 0{,}193$). E, se aplicarmos teste *t* para amostras pareadas, obtém-se valor bicaudal $\rho = 0{,}061$. Finalmente, se aplicarmos o teste qui-quadrado, vamos obter valor de cauda $\rho = 0{,}675$.

Como se verificou, todos os testes conduziram a conclusão que contraria o que se esperava. Isso ocorreu porque, para que um teste estatístico funcione adequadamente para o problema em questão, a respectiva variável de teste X_{cal}, calculada em função dos dados das duas amostras, a ser utilizada para determinar $\rho = P[X \geq X_{cal}]$, deve possuir três propriedades: a) considerar a amplitude da diferença observada em cada par de valores relacionados a cada classe da variável aleatória; b) acumular as diferenças em sentidos opostos observadas em classes distintas (impedir que uma anule a outra); e c) ajustar-se a uma distribuição de probabilidade conhecida *X*.

O único teste, dentre os aplicados, que apresenta as duas primeiras propriedades é o qui-quadrado mas, para atender à terceira, é necessário que pelo menos 80% das células possuam frequência maior do que 5 e que nenhuma célula apresente frequência menor do que 1 (Siegel & Castellan, 2006), o que não ocorre com os dados da Tabela 2.

O qui-quadrado também pode falhar se os valores contidos nas células são esparsos ou possuem forte desequilíbrio (ver exemplo na seção 7).

Como alternativa ao teste qui-quadrado, quando as condições anteriores não são atendidas, surgem os testes exatos, sendo o teste Fisher, proposto em 1925 (Fisher, 1970), o primeiro deles, o qual é aplicável a duas pequenas amostras de variáveis com duas categorias (tabelas com $l = 2$ linhas e $c = 2$ colunas). Esse teste foi mais tarde estendido para tabelas com $l > 2$ e $c > 2$ por Freeman & Halton (1951). Porém sua aplicação exige grande esforço computacional, principalmente se o número de classes for grande (Sprent & Smeeton, 2000, p. 322). Nesses casos, deve-se dispor de *softwares* apropriados como, por exemplo, o StatXact (2008).

A insegurança na utilização do qui-quadrado em problemas com pequenas amostras e a dificuldade de aplicação dos testes exatos levaram os autores a propor um novo teste não paramétrio para abordar problemas com pequenas amostras de variáveis categorizadas e a realizar estudos comparativos sobre o desempenho desses três testes, ou seja, sobre a capacidade em decidir corretamente sobre a hipótese H_0.

Na seção a seguir, apresenta-se a teoria dos testes exatos, com destaque para o teste de Fisher, e o procedimento adotado pelo *software* o StatXact para essa classe de problemas, cujo principal objetivo é mostrar a dificuldade de solução de problemas de pequenas amostras cujas variáveis assumem mais de duas categorias.

4 Testes exatos baseados na teoria das permutações

Para exemplificar a aplicação do teste exato de Fisher às tabelas de dimensão 2 × 2, considere as Tabelas 3a-c, nas quais o Grupo I refere-se ao sexo masculino e o Grupo II, ao sexo feminino.

Na linha superior de cada uma dessas tabelas estão as frequências de pessoas com altura igual ou superior a 1,80 metro e, na linha inferior, as frequências de pessoas com altura inferior a 1,80 metro, obtidas de uma amostra de 8 homens e 9 mulheres. Deseja-se verificar, com base nessa pequena amostra, se homens possuem estatura superior à das mulheres. Considere que a hipótese H_0 estabelece a igualdade das alturas e a hipótese alternativa H_1, que a altura dos homens é superior à das mulheres. Para aplicar o teste exato de Fisher sobre esse problema, determina-se o valor de $\rho = \sum_{i=1}^{r} p_i$, onde p_i é a probabilidade de ocorrer uma situação igual ou mais extrema (no sentido da

Tabela 2. Dados para aplicação dos testes disponíveis na literatura.

Amostra	*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

Tabela 3. Dados para exemplificação do teste exato de Fisher.

Grupos			Grupos			Grupos		
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)	

Fonte: Autores.

hipótese H_1) do que a da Tabela 3a, mantendo-se fixos os valores totais marginais. Observe que a amostra forneceu 6 homens com estatura superior e 2 com estatura inferior a 1,80 m. Como o teste é unilateral (devido à hipótese alternativa H_1), existem duas outras situações mais extremas do que a da Tabela 3a com valores marginais fixos, as quais estão representadas pelas Tabelas 3b e 3c.

A probabilidade exata de se observar um conjunto particular de frequências em uma Tabela 2 × 2, quando os totais marginais são considerados como fixos, é dada pela distribuição hipergeométrica, resultando ρ = 0,109, obtido da soma das parcelas $p_{(a)}$, $p_{(b)}$ e $p_{(c)}$, dadas pela Equações 2, 3 e 4, respectivamente.

$$p_{(a)} = \frac{9!\ 8!\ 8!\ 9!}{17!\ 6!\ 3!\ 2!\ 6!} = 0,0968 \quad (2)$$

$$p_{(b)} = \frac{9!\ 8!\ 8!\ 9!}{17!\ 7!\ 2!\ 1!\ 7!} = 0,0012 \quad (3)$$

$$p_{(c)} = \frac{9!\ 8!\ 8!\ 9!}{17!\ 8!\ 1!\ 0!\ 8!} = 0,0004 \quad (4)$$

Neste caso, como ρ > 0,05, não é possível rejeitar H_0 com certeza de 95%.

A seguir será apresentado um exemplo para ilustrar como é aplicado o teste exato para tabelas com $l > 2$ e $c > 2$.

Considere os dados da Tabela 4 representando o número de executivos pertencentes a quatro unidades de negócio de uma grande corporação que obtiveram avaliações alta, média e baixa em um programa de promoção de executivos. Com base nessa pequena amostra é possível concluir que a unidade de negócio *A* possui executivos mais capazes (hipótese alternativa H_1)?

Se fosse aplicado o teste qui-quadradro, a estatística construída teria $(l-1)\times(c-1) = 6$ graus de liberdade e forneceria $\chi^2 = 11,555$. Como $P(\chi^2_6 > 11,555) = 0,0726$, não seria possível rejeitar a hipótese nula H_0 com certeza de 95% e afirmar que a unidade de negócio *A* possui executivos mais capazes. Para uma discussão sobre o teste qui-quadrado ver, por exemplo, Siegel & Castellan (2006).

Para aplicar o teste exato são geradas todas as possíveis tabelas a partir da configuração dos dados da amostra, mantendo-se fixos os valores marginais. Aquelas tabelas que originarem valores de $\chi^2 \geq 11,555$ representam situações mais estremas que a da amostra original e portanto contribuem com seus respectivos valores de *p* para compor o valor de ρ. Por exemplo, as Tabelas 5a e 5b são dois possíveis arranjos obtidos da Tabela 4. A primeira fornece $\chi^2 = 14,676$, e deve ser considerada como uma situação mais extrema do que a da amostra original. Assim, seu respectivo valor de *p* contribui na determinação de ρ. Já a Tabela 5b fornece $\chi^2 = 9,778$ e seu correspondente valor de *p* não contribui para o cálculo de ρ.

Tabela 4. Resultado da avaliação de executivos.

Nível de avaliação	Unidades de negócio				Totais
	A	B	C	D	
Alto	5	2	2	0	9
Médio	0	1	0	1	2
Baixo	0	2	3	4	9
Totais	5	5	5	5	20

Fonte: Autores.

A generalização do cálculo da probabilidade *p* de um conjunto particular de frequências para uma tabela com *l* linhas e *c* colunas feita por Freeman & Halton (1951) é dada pela Equação 5, na qual $n_{i,o}$ é o valor marginal da linha *i*, $n_{o,j}$ é o valor marginal da coluna *j*, n_{ij} é o valor contido na célula (i, j) e *n* é a soma dos valores de todas as células:

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

Na aplicação do teste exato a tabelas de dimensão $l\times c$, todas as possíveis tabelas oriundas dos dados originários da amostra devem ser representadas e é a representação dessas tabelas que, em geral, requer grande esforço computacional.

Esse tipo de problema pode ser resolvido, por exemplo, pelo software StatXact (2008) que, para esse particular caso, fornece ρ = 0,0398 o que, em contradição ao resultado do teste qui-quadrado, leva a rejeitar a hipótese nula H_0 com certeza de 95%.

5 Teste baseado na diferença de duas distribuições uniformes

Nesta seção é apresentado um novo teste não paramétrico para o problema em questão, cuja estatística de teste é dada pela diferença de duas distribuições uniformes de probabilidades.

Tabela 5. Duas permutações dos resultados da avaliação de executivos.

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)		

Sejam $j = 1, 2, ..., k, k \geq m$ os índices das alternativas que uma variável aleatória categorizada C pode assumir e sejam $P = \{p_j, j = 1, 2, ..., k\}$ e $Q = \{q_j, j = 1, 2, ..., k\}$ as verdadeiras distribuições de probabilidades dessa variável em duas populações distintas P_1 e P_2 (por exemplo, empresas mais competitivas e menos competitivas). Considere as funções dadas pelas Equações 6 e 7.

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

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

Então, se $p_j = q_j$, para todo $j = 1, 2, ..., k$ pode-se verificar facilmente que $p_j^{'} = q_j^{'} = 1/k$, para todo j. Ou seja, se P e Q possuem mesma distribuição de probabilidades, então as funções p_j e q_j convertem a distribuição das estratégias j, para ambas as populações de empresas, em uma distribuição uniforme com probabilidade igual a $1/k$ para todo j. Isso mostra que o teste proposto, que na sua essência baseia-se na verificação da diferença $|p_j^{'} - q_j^{'}|$, é convergente.

Sejam agora f_j e $g_j, j = 1, 2, ..., m,$ as frequências que a variável aleatória C assume em duas amostras A_1 e A_2 de tamanhos n_1 e n_2 extraídas das populações P_1 e P_2, respectivamente. Como f_j / n_1 e g_j / n_2 são estimativas justas para p_j e q_j, respectivamente, se A_1 e A_2 forem amostras de uma mesma população, então as Equações 8 e 9 devem possuir valores próximos de $1/m$, para todo $j = 1, 2, ..., m$, para quaisquer valores de n_1 e n_2. Foi esse fato que motivou a proposição desse teste para o caso de amostras pequenas, apesar de se tratar de um teste assintótico.

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

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

Considere agora a estatística $D = \sum_{j=1}^{m} |u_j - v_j|$ onde u_j e v_j são frequências relativas da variável $j = 1, 2, ..., m,$ tal que $\Pr(j) = 1/m$, para todo j. Essa variável é pouco sensível à variação do número de elementos da amostra (pelo menos para pequenas variações, o que sempre ocorre quando se trata de amostras de pequeno tamanho, como no caso em questão), mas ela depende do valor de m, uma vez que é proveniente da soma de m parcelas, cada uma dada pela diferença de duas variáveis uniformes. A distribuição de probabilidades dessa estatística não é conhecida. Porém é possível, por meio de simulação, construir seu histograma para diversos valores de m e, a partir de cada um desses histogramas, determinar D_α, onde D_α é o valor de D que deixa α% dos dados à sua direita.

Com auxílio dessa informação é possível verificar se as listas de estratégias A_1 e A_2 provêm de uma mesma população (hipótese H_0). Basta calcular a estatística $D_{cal} = \sum_{j=1}^{m} |r_j - s_j|$ a partir dos valores de f_j e g_j originados de A_1 e A_2, respectivamente, e confrontar com o valor de D_α. Se $D_{cal} > D_\alpha$, podendo-se rejeitar a hipótese H_0 com nível de certeza (1-α).

Observe-se que a variável D_{cal} (assim como D) é definida no intervalo [0, 2]. Quando $f_j = g_j$, para todo $j = 1, 2, ..., m$, então $D_{cal} = 0$, o que fornece a máxima certeza de que ambos os conjuntos A_1 e A_2 são provenientes de uma mesma população. Agora, quando, para cada $j = 1, 2, ..., m$, ($f_j = 0$, $g_j > 0$) ou ($f_j > 0$, $g_j = 0$), o que significa que cada grupo de empresas declarou conjuntos distintos de estratégia e portanto a interseção dos conjuntos A_1 e A_2 é vazia, então $D_{cal} = 2$, o que fornece a máxima certeza de rejeição da hipótese nula H_0.

5.1 Determinação do valor de D_α

A determinação de D_α foi feita a partir do histograma da variável D, construído por meio de um processo de simulação computacional, procedimento fornecido a seguir ilustrado para o caso de $m = 6$, $n_1 = n_2 = 12$.

Passo 1. Estabelecer a seguinte correlação, conforme Tabela 6, em que *NA* é um número aleatório retangular no intervalo [0, 1].

Passo 2. Gerar n_1 números aleatórios retangulares (*NA*) no intervalo [0, 1] para a primeira amostra e outros n_2 números para a segunda amostra e obter os conjuntos A_1 e A_2, ou seja, valores de f_j e g_j. Para $n_1 = n_2 = 12$, um possível resultado é mostrado nas colunas f_j e g_j da Tabela 7 onde, dentre os 12 valores sorteados para a amostra A_1, 2 deles caíram no intervalo [0, 1/6), e para a amostra A_2, 3 valores caíram nesse mesmo intervalo, originando então $f_1 = 2$ e $g_1 = 3$.

Passo 3. Determinar, para cada amostra gerada A_1 e A_2, $D = \sum_{j=1}^{m} |u_j - v_j|$, em que $u_j = (f_j / n_1)$, $v_j = (g_j / n_2)$, conforme mostra a Tabela 7, que fornece, para esse exemplo, $D = 0{,}333$.

Passo 4. Repetir 10.000 vezes os passos 1 a 3, gerando 10.000 valores ordenados para D, e identificar o valor de D_α para os níveis de significância $\alpha = 0{,}01$ e $\alpha = 0{,}05$ ($D_{0,05}$ é dado pelo valor de D que deixa 500 valores à sua direita e $D_{0,01}$ é dado pelo valor de D que deixa 100 valores à sua direita). A Tabela 8 fornece os valores críticos de D_α para diversos valores de m, $\alpha = 0{,}05$ e $\alpha = 0{,}01$.

Aplicando-se o teste para os dados da Tabela 1, obtém-se $D_{cal} = 0{,}493$. Como nesse exemplo $m = 6$,

Tabela 6. Relação entre número aleatório retangular e as classes da variável.

NA no intervalo	Variável *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

Tabela 7. Aplicação do teste de diferença de duas distribuições uniformes.

Estratégia (*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
Soma	12	12	1,000	1,000	0,333

Fonte: Autores.

Tabela 8. Valores críticos de 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

Fonte: Autores.

conclui-se que não se pode rejeitar a hipótese nula H_0 e deve-se aceitar que os dois grupos de empresas adotam conjuntos semelhantes de estratégia.

6 Estudo sobre o poder dos testes

A eficácia dos testes exato, qui-quadrado e proposto foi avaliada por meio da análise da curva de poder, que fornece a probabilidade de aceitação (*Pa*) da hipótese nula (H_0) em função do nível de semelhança entre as duas amostras.

A curva de poder foi levantada por meio de simulação computacional em função do nível de semelhança entre as amostras, definido pelo parâmetro denominado grau de simetria (GS) das distribuições das amostras A_1 e A_2, variando no intervalo [0, 1] e dado pela Equação 10.

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

em que p_j e q_j são as probabilidades de a variável categorizada originada das amostras A_1 e A_2 para todo $j = \{1, 2, ..., m\}$.

Definindo-se valores apropriados para *pj* e *qj*, foram obtidas, por simulação, amostras provenientes de populações com os seguintes graus de simetria *GS* = {0,0; 0,2; 0,4; 0,6 e 0,8}. Observe que se *pj* = *qj* para todo *j*, a Equação 2 fornece *GS* = 0 e as amostras obtidas por simulação para esse caso serão provenientes de uma mesma população. Por outro lado, se *pj* = 0 quando *qj* ≠ 0, para todo *j,* então *GS* = 1, o que origina configurações com amostras provenientes de populações totalmente distintas.

Foram feitos ensaios computacionais para as seis seguintes configurações de problemas identificadas pelos conjuntos de valores de (m, n_1, n_2): (3, 7, 7), (4, 8, 8), (5, 10, 10), (6, 12, 12), (7, 14, 14) e (8, 16, 16). Para cada um desses seis casos e para cada um dos cinco valores de *GS* anteriormente citados, determinou-se a probabilidade de aceitação *Pa* segundo o teste exato, o qui-quadrado e o teste proposto.

Para isso, foram gerados 100 problemas para cada um dos seis conjuntos de valores (m, n_1, n_2) e cada um dos cinco níveis de semelhança *GS*. O valor de *Pa* para um determinado teste e para um dado conjunto de valores (m, n_1, n_2) e um dado valor de *GS* puderam então ser identificados pela contagem direta do número de problemas em que ocorria a aceitação de H_0.

Adotou-se, para todos os testes, nível de significância α = 0,05. Assim a aceitação de H_0 ocorria sempre que ρ = P[$X > X_{cal}$] > α, em que *X* é a variável do teste e X_{cal} é o valor da estatística do teste, ou sempre que $X_{cal} < X_{crit}$, em que X_{crit} é tal que P[$X > X_{crit}$] = α, o que é a mesma coisa vista de duas maneiras.

No total, foram ensaiados, portanto, 3.000 problemas, 100 para cada combinação [(m, n_1, n_2); *GS*], e cada um deles foi resolvido pelos três testes.

A configuração de cada problema, ou seja, valores de *fj* e *gj*, para ambas as amostras, foi obtida de forma análoga àquela descrita nos passos 1 e 2 do procedimento para determinação de D_α, apresentado na seção 5.

Desta curva, levantada por meio de simulação computacional, puderam ser extraídos os seguintes indicadores para análise comparativa dos testes:

a) Risco α, que é a probabilidade de se cometer o erro tipo I (rejeitar a hipótese nula quando ela é verdadeira), dado por α = (1-*Pa*), para *GS* = 0;

b) Média dos riscos β dado pela média de *Pa* para os quatro valores de *GS* > 0, onde β é a probabilidade de se cometer o erro tipo II (aceitar a hipótese nula quando ela é falsa); e

c) Indicador característico da curva de poder (*IC*), determinado pela relação $(Declividade)_{0,50}/(GS)_{0,50}$, onde $(Declividade)_{0,50}$ é a inclinação da curva no ponto $(GS)_{0,50}$, sendo $(GS)_{0,50}$ o valor de *GS* que origina uma probabilidade de aceitação de 50%.

O valor de $(Declividade)_{0,50}$ foi determinado pela Equação 11:

$$(Declividade)_{0,50} = -\frac{(GS)_{0,6} - (GS)_{0,4}}{100.(0,6-0,4)} \quad (11)$$

Como a curva é decrescente, introduziu-se o sinal negativo para tornar o resultado da declividade positivo. Multiplicou-se o denominador por 100 para representá-la em uma escala mais adequada (intervalo 1 a 10). Os valores de $(GS)_{0,40}$ e $(GS)_{0,60}$ foram obtidos por inspeção visual do gráfico da curva de poder gerada pelos cinco pontos (*GS*, *Pa*).

Os dois parâmetros $(Declividade)_{0,50}$ e $(GS)_{0,50}$ são muito utilizados para se avaliar o poder discriminante de planos de inspeção de qualidade. Quanto maior o valor de $(Declividade)_{0,50}$ e quanto menor o valor de $(GS)_{0,50}$, maior o poder do plano, ou o poder do teste estatístico, no presente estudo. Assim, o índice *IC* expressa em um só indicador as propriedades de ambos (quanto maior seu valor, maior o poder do teste) e pode dirimir dúvidas que porventura restem da aplicação dos indicadores de risco α e β.

Estudos sobre desempenho de testes estatísticos adotam apenas os indicadores de risco, conforme foi feito, por exemplo, por Tanizaki (1997). Assim, a utilização de um novo indicador (*IC*) com a propriedade acima citada traz alguma contribuição para esse tipo de estudo.

As Tabelas 9a a 9f fornecem os resultados obtidos dos ensaios realizados com os testes exato (solução obtida pelo StatXact), qui-quadrado (*Q-Q*) e uniforme. Os valores de *Pa* estão expressos em porcentagem, pois correspondem diretamente ao número de problemas em que ocorreu a aceitação de H_0, em 100 problemas ensaiados para cada valor de *GS*. O significado e a forma de obtenção dos valores de *IC*, α, e β médio que aparecem nas Tabelas 9a-f serão explicados na seção seguinte.

Tabela 9a. Resultados para $m = 3$, $n_1 = n_2 = 7$.

Teste	Probabilidade de aceitação (*Pa*) - Porcentagem					*IC* e riscos (%)		
	***GS* = 0**	***GS* = 0,2**	***GS* = 0,4**	***GS* = 0,6**	***GS* = 0,8**	**IC**	**α**	**β médio**
Exato	98	94	80	24	0	5,6	2	50
Q-Q	95	89	73	16	0	5,9	5	45
Uniforme	97	92	77	20	0	6,6	3	47

Fonte: Autores.

Tabela 9b. Resultados para $m = 4$, $n_1 = n_2 = 8$.

Teste	Probabilidade de aceitação (*Pa*) - Porcentagem					*IC* e riscos (%)		
	***GS* = 0**	***GS* = 0,2**	***GS* = 0,4**	***GS* = 0,6**	***GS* = 0,8**	**IC**	**α**	**β médio**
Exato	97	94	78	51	12	2,3	3	59
Q-Q	96	93	78	50	11	2,3	4	58
Uniforme	96	93	82	58	17	1,3	4	63

Fonte: Autores.

Tabela 9c. Resultados para $m = 5$, $n_1 = n_2 = 10$.

Teste	Probabilidade de aceitação (*Pa*) - Porcentagem					*IC* e riscos (%)		
	***GS* = 0**	***GS* = 0,2**	***GS* = 0,4**	***GS* = 0,6**	***GS* = 0,8**	**IC**	**α**	**β médio**
Exato	96	92	78	34	3	4,2	4	52
Q-Q	97	94	83	38	5	4,1	3	55
Uniforme	95	92	84	39	4	3,4	5	55

Fonte: Autores.

Tabela 9d. Resultados para $m = 6$, $n_1 = n_2 = 12$;

Teste	Probabilidade de aceitação (*Pa*) - Porcentagem					*IC* e riscos (%)		
	***GS* = 0**	***GS* = 0,2**	***GS* = 0,4**	***GS* = 0,6**	***GS* = 0,8**	**IC**	**α**	**β médio**
Exato	90	94	71	28	6	4,3	10	50
Q-Q	92	96	77	38	7	3,5	8	55
Uniforme	91	85	74	39	9	3,9	9	52

Fonte: Autores.

Tabela 9e. Resultados para $m = 7$, $n_1 = n_2 = 14$.

Teste	Probabilidade de aceitação (*Pa*) - Porcentagem					*IC* e riscos (%)		
	***GS* = 0**	***GS* = 0,2**	***GS* = 0,4**	***GS* = 0,6**	***GS* = 0,8**	**IC**	**α**	**β médio**
Exato	90	94	67	32	3	3,5	10	49
Q-Q	95	95	73	42	3	2,7	5	53
Uniforme	94	96	70	46	3	2,3	6	54

Fonte: Autores.

Tabela 9f. Resultados para $m = 8$, $n_1 = n_2 = 16$.

Teste	Probabilidade de aceitação (*Pa*) - Porcentagem					*IC* e riscos (%)		
	***GS* = 0**	***GS* = 0,2**	***GS* = 0,4**	***GS* = 0,6**	***GS* = 0,8**	**IC**	**α**	**β médio**
Exato	95	88	68	21	1	4,9	5	45
Q-Q	97	94	79	26	2	5,8	3	50
Uniforme	91	91	72	32	5	3,9	9	50

Fonte: Autores.

7 Análise dos resultados e conclusões

A eficácia dos testes foi avaliada pelos riscos α e β e pelo indicador característico da curva de poder (*IC*).

O risco α para cada configuração de problema (m, n_1, n_2) é dado, em porcentagem, na Tabela 9, pelo valor (100- *Pa*) para a coluna $GS = 0$, uma vez que o valor de *Pa* corresponde, entre os 100 ensaios realizados, à quantidade deles em que o teste conduziu à decisão acertada, ou seja, aceitar a hipótese H_0 quando ela é verdadeira. Já o risco β, também em porcentagem, é dado pela média dos valores de *Pa*, para todo $GS = \{0,2, 0,4\ 0,6, 0,8\}$, ou seja, probabilidade de aceitar H_0 quando ela não é verdadeira (amostra apresenta grau de simetria diferente de zero).

Os valores de $(Declividade)_{0,50}$, para cada configuração (m, n_1, n_2), foram calculados pela Equação 3. Esses três parâmetros de análise estão apresentados na Tabela 10.

Analisando-se os riscos α e β dados na Tabela 10, verifica-se que o teste qui-quadrado é o que apresenta menor risco α dos três e risco β intermediário ao dos outros dois, mas, com relação ao indicador *IC*, é ele que apresenta o menor desempenho dos três.

O teste proposto apresenta riscos α e β, assim como indicador característico (*IC*) parecidos com os do teste exato, o que evidencia que ambos possuem desempenho bastante similar.

A Tabela 11 apresenta a quantidade de problemas que cada teste decidiu de forma acertada, dentre os 3.000 problemas ensaiados. Verifica-se que o teste exato foi o que mais acertou na decisão (1.753 vezes), enquanto que o teste proposto apresentou um desempenho um pouco inferior aos outros dois.

Essa análise nos permite concluir que os testes exato e proposto apresentam desempenhos bastante próximos, o que parece correto, e que o teste qui-quadrado supera ambos, pelo menos como instrumento de decisão quando a hipótese nula é verdadeira. Essa é uma conclusão até certo ponto inesperada, em se tratando de problemas com pequenas amostras. Seria então o teste qui-quadrado uma alternativa válida para o teste exato?

Considerando o exemplo dos dados da Tabela 12, verifica-se que nem sempre. Aplicando-se o teste

Tabela 10. Resumo dos parâmetros de avaliação da eficácia dos testes.

Parâmetro	*m*	Teste		
		Exato	**Q quad**	**Uniforme**
Risco 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
	Valor médio	5,7	4,7	6,0
Risco β médio (%)	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
	Valor médio	50,5	52,6	53,3
IC	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
	Valor médio	3,9	3,1	4,1

Fonte: Autores.

Tabela 11. Número de problemas com decisão acertada.

GS	Testes		
	Exato	**Q-Q**	**Uniforme**
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
Todos	1753	1710	1684

Fonte: Autores.

Tabela 12. Exemplo de problema com três amostras.

Amostra	**Valores**								
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

Fonte: StatXact (2003).

exato aos dados dessa tabela obtém-se, pelo StatXact, $\rho = 0{,}0013$, o que evidencia que as três amostras não provêm de uma mesma população. O teste qui-quadrado, por sua vez, fornece valor de $\rho = 0{,}1342$, mostrando claramente que para amostras pequenas com dados que apresentam forte desbalanceamento, como é o caso desse exemplo, esse teste não funciona bem. E a Tabela 4 fornece outro exemplo desse fenômeno. Assim, sua utilização generalizada leva a decisões não confiáveis, daí a necessidade de se buscar testes alternativos.

E o teste aqui proposto, como se comporta frente a esse tipo de amostra?

Para responder a essa questão, inicialmente, é preciso observar que, embora o teste proposto tenha sido direcionado a problemas com duas amostras, é possível resolver também problemas com mais amostras, bastando aplicá-lo às diversas combinações de amostras, duas a duas.

Aplicando-se o teste uniforme aos dados da Tabela 12, duas a duas (observe que é necessário eliminar as colunas que contêm zeros em ambas as amostras), obtêm-se valores para D_{cal} iguais a 1,959, 1,622 e 1,964 para as combinações *A*/*B*, *A*/*C* e *B*/*C* de amostras, respectivamente. Como o máximo valor de *Dcal* é 2,0, o teste indica com alto nível de certeza que a amostra *B* provém de uma população distinta das demais, o que o qui-quadrado não conseguiu identificar.

Se aplicarmos agora o teste proposto aos dados da Tabela 4, obteremos valores de D_{cal} iguais a 1,750, 1,556 e 2,000 para as amostras *A*/*B*, *A*/*C* e *A*/*D*, respectivamente. Como $D_{\alpha = 0{,}01} = 1{,}429$, para $m = 3$ (caso da Tabela 4) conclui-se, com alto grau de certeza, que a unidade de negócio *A* possui executivos mais capazes.

Esses dois exemplos mostram que a melhor alternativa ao teste exato, que apresenta sérias dificuldades de aplicação, é o teste proposto e não o qui-quadrado que, embora tenha demonstrado melhor desempenho no conjunto dos ensaios, pode falhar conforme a instância do problema.

Referências

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.