AVALIAÇÃO MONTE CARLO DO TESTE DE NORMALIDADE DE QUI-QUADRADO SOB DIFERENTES CRITÉRIOS DO NÚMERO DE CLASSES

Tales Jesus FERNANDES[1]
Eric Batista FERREIRA[2]
Daniel Furtado FERREIRA[3]

▪ RESUMO: Sabe-se que a densidade de distribuições de probabilidade pode ser aproximada pelo histograma amostral. A densidade normal é a mais utilizada, pois essa suposição existe na maioria dos testes estatísticos. Uma das formas de verificar a normalidade dos dados é através do teste de aderência de qui-quadrado. No entanto, quando este teste é utilizado, os autores geralmente não explicitam qual foi a regra usada para determinar o número de classes do histograma (que é o primeiro passo na execução deste teste). Dessa forma, um possível mau desempenho é atribuído ao teste qui-quadrado, mas suspeita-se que ele possa ser agravado pela escolha de um critério não ótimo de determinação do número de classes (k). Os objetivos deste trabalho foram: verificar a validade dessa hipótese e apontar qual o critério mais aconselhado. Comparou-se o desempenho (taxa de erro tipo I e poder) de cinco critérios de determinação de k (Critério empírico, Sturges, Scott, Doane e Freedman e Diaconis) no teste de normalidade de qui-quadrado, via simulação Monte Carlo. Concluiu-se que os resultados do teste qui-quadrado são afetados pela escolha do valor de k e o critério que melhor controlou as taxas de erro tipo I foi o critério de Sturges. O critério de Freedman e Diaconis apresentou os maiores valores de poder.

▪ PALAVRAS-CHAVE: Testes de aderência; erro tipo I; poder; histograma.

[1]Universidade Federal de Lavras – UFLA, Programa de Pós-graduação em Estatística e Experimentação Agropecuária, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *talesest@yahoo.com.br*
[2]Universidade Federal de Alfenas – UNIFAL, Instituto de Ciências Exatas, CEP: 37130-000, Alfenas, MG, Brasil. E-mail: *eric.ferreira@unifal-mg.edu.br*
[3]Universidade Federal de Lavras – UFLA, Departamento de Ciências Exatas, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *danielff@dex.ufla.br*

1 Introdução

A suposição de normalidade dos dados amostrais ou experimentais é uma condição exigida para a realização de muitas inferências válidas a respeito de parâmetros populacionais. Vários dos diferentes métodos de estimação e testes de hipóteses existentes foram formulados sob a suposição de que a amostra aleatória tenha sido extraída de uma população normal. A rejeição desse pressuposto deve conduzir os pesquisadores ao uso de técnicas não paramétricas (CAMPOS, 1983), ou aos modelos lineares generalizados (MCCULLAGH e NELDER, 1989), ou ainda ao uso de técnicas de estatística computacional intensivas. Muitas vezes essa rejeição se deve a baixa qualidade do critério de determinação do histograma amostral como estimador da densidade normal.

A densidade de diversas distribuições de probabilidade pode ser aproximada pelo histograma amostral. Nessa situação, busca-se em geral minimizar a soma de quadrados dos desvios entre o histograma e a densidade (BROWN e HWANG, 1993). Diversos pesquisadores têm demonstrado particular interesse pela aproximação de densidades ao longo da história (DOANE, 1976; SCOTT, 1979; FREEDMAN e DIACONIS, 1981; BROWN e HWANG, 1993). Entre as densidades mais usadas, destaca-se a distribuição normal (JOHNSON e KOTZ, 1970a,b). A identificação de normalidade dos dados é feita através de testes estatísticos, sendo que entre os testes mais comuns e mais encontrados na literatura, estão: o teste de aderência qui-quadrado, Kolmogorov-Smirnov, Lilliefors e Shapiro-Wilk.

Conforme é dito por diversos autores, tais como, Cirilo e Ferreira (2003), Mendes e Pala (2003), Seier (2004), Silva e Ferreira (2008) e Torman *et al.* (2012), o teste Shapiro-Wilk é o que melhor controla a taxa de erro tipo I e apresenta maior poder, sendo que dentre os testes citados acima ainda em concordância com estes autores, o teste que apresenta pior desempenho é o qui-quadradado. Um teste que controla a taxa de erro tipo I com taxas iguais aos níveis nominais fixados é chamado "teste exato". Um teste que controla as taxas de erro tipo I com taxas inferiores ao nível de significância adotado é denominado "teste conservativo", e se o teste controla as taxas de erro tipo I com níveis superiores a significância adotada este é chamado de "teste liberal".

Neste trabalho foi abordado o teste de normalidade de qui-quadrado devido a sua versatilidade, isto é, este é um teste que se presta não apenas para verificar a normalidade dos dados, mas sua aderência a qualquer distribuição de frequências, além de poder ser utilizado como teste de independência entre variáveis. E ainda este teste é o mais ensinado nos cursos de graduação devido à sua facilidade de execução e compreensão por parte dos alunos. Inclusive, é fácil encontrar exemplos de aplicação do teste de qui-quadrado em diversas áreas do conhecimento, destacando-se a biociências (NEVES *et al.*, 2002; CABRAL, 2007; PEDROSO *et al.*, 2010; MOTA *et al.*, 2012; PINHAT *et al.*, 2012; YEH *et al.*, 2012). Este teste possui algumas limitações como, por exemplo, ser afetado por tamanhos amostrais pequenos, classes com frequências esperadas inferiores a 1 (um) e pelo número de classes gerados por diferentes algoritmos.

Mas o teste qui-quadrado é geralmente utilizado sem descrever como se obteve o número de classes k (LOPES, 2005; CIELASK *et al.*, 2007; ROSASCO, 2009; LUNZ *et al.*, 2010; CAMPOS *et al.*, 2012; LINHARES *et al.*, 2012; MORAES *et al.*, 2012; SARCINELLI *et al.*, 2012). Com isso, ao demonstrar um desempenho abaixo do esperado, o teste de aderência de qui-quadrado é taxado como não ótimo. Entretanto, uma das razões para esse baixo desempenho pode ser a forma com que o número de classes foi determinado. Em outras palavras, acredita-se que o baixo desempenho apresentado pelo teste de qui-quadrado pode ser agravado pela escolha do valor de k. Não se espera que o teste qui-quadrado supere os demais testes na detecção de normalidade dos dados, pois estes são testes específicos de normalidade, mas a opção pelo melhor critério a ser utilizado no primeiro passo deste teste pode torná-lo mais competitivo perante os demais, proporcionando-lhe melhores resultados.

Os passos de um teste qui-quadrado para distribuições contínuas são:

1. Determinar o número de classes *(k)* ou intervalo em que a amostra será dividida.
2. Verificar as frequências observadas (FO) em cada classe.
3. Obter a frequência esperada pela distribuição contínua em questão (FE).
4. Com as frequências observadas e as esperadas calcular a estatística de teste qui-quadrado:
$$\chi_c^2 = \sum_{i=1}^{k} \frac{(FO_i - FE_i)^2}{FE_i}. \quad (1)$$
5. Comparar o valor obtido no 4° passo com o quantil de qui-quadrado tabelado, adotando-se α de 5% e 1%, e graus de liberdade $k-1$. As hipóteses em questão são:
$$\begin{cases} H_0: & \text{Os dados seguem uma distribuição normal;} \\ H_1: & \text{Os dados não seguem uma distribuição normal.} \end{cases}$$

Existem diferentes maneiras de se fazer o primeiro passo deste teste:

Segundo Sturges (1926), a idéia básica dos trabalhos de ajuste da distribuição normal pelos histogramas amostrais, é de que uma variável normal pode ser apropriadamente dividida, de tal forma que as frequências das classes abranjam uma série binomial de potência 2. Ainda de acordo com este autor, o número ótimo de classes (k) é, em geral:
$$k = 1 + \log_2(n), \quad (2)$$
em que n é o tamanho da amostra.

Doane (1976) critica esta regra pelo fato de não fornecer um número de classes grande o suficiente para revelar a forma de distribuições severamente assimétricas e propõe a adição de classes extras (k_e), dadas por:
$$k_e = \log\left(1 + \frac{|\sqrt{b_1}|}{S_{\sqrt{b_1}}}\right),$$

em que, $S_{\sqrt{b_1}}$ é o erro padrão do coeficiente de assimetria $\sqrt{b_1}$, os quais são dados por:

$$\sqrt{b_1} = \frac{\sum_{i=1}^{n}(x-\overline{x})^3}{\left[\sum_{i=1}^{n}(x-\overline{x})^2\right]^{\frac{3}{2}}}$$

$$S_{\sqrt{b_1}} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$

Daí a regra de números de classes proposta por Doane (1976) fica:

$$k = 1 + \log_2(n) + k_e = 1 + \log_2(n) + \log\left(1 + \frac{|\sqrt{b_1}|}{S_{\sqrt{b_1}}}\right). \tag{3}$$

Critérios empíricos têm sido usados em livros textos de estatística (FERREIRA, 2005). Um exemplo de critério empírico será considerado neste trabalho:

$$k = \begin{cases} \sqrt{n}, & \text{se } n \leq 100; \\ 5\log_{10}(n), & \text{se } n > 100. \end{cases} \tag{4}$$

Esse critério só depende do tamanho amostral, portanto, para diferentes naturezas da densidade de uma variável aleatória para amostras de mesmo tamanho, o número calculado de classes será o mesmo.

Scott (1979) propõe que o critério do número de classes *(k)*, baseado em uma população normal, seja:

$$k = 1 + \frac{A\sqrt[3]{n}}{3,49S}, \tag{5}$$

sendo A a amplitude amostral total e S o desvio padrão amostral.

Segundo Scott (1979), esse critério é útil para uma série de densidades, considerando-se as Gaussianas e não-Gaussianas, e que conduz a diferentes números ótimos de classes. Densidades assimétricas, como a log-normal; distribuição de t de Student, de caudas pesadas; e, finalmente, uma distribuição bimodal, dada por uma mistura de duas normais, foram usadas para avaliar teoricamente a validade do critério proposto.

O software estatístico R (R Development Core Team, 2012), em sua função nclass.FD(), calcula o número de classes da seguinte forma:

$$k = \begin{cases} mad(x), & \text{se } h = 0; \\ c\left(\frac{A}{2hn^{-\frac{1}{3}}}\right), & \text{se } h > 0; \end{cases} \tag{6}$$

sendo $h = q_{75\%}(x) - q_{25\%}(x)$, em que $q_{\gamma\%}(x)$ é o quantil amostral de $\gamma_{\%}$, *mad(x)* calcula o desvio absoluto mediano, ou seja, a mediana dos valores absolutos dos desvios em relação a mediana, e *c(x)* é o primeiro inteiro maior que *x*.

O objetivo do presente trabalho foi verificar se os resultados do teste qui-quadrado são diretamente afetados pela determinação do valor de k e, se forem,

indicar qual o melhor critério entre os citados anteriormente. Estes critérios foram comparados quanto à acurácia do ajuste a densidade normal por meio do teste de qui-quadrado, via simulação Monte Carlo. Tal comparação foi capaz de detectar qual dessas construções mais bem controla as taxas de erro tipo I e apresenta maior poder de acordo com este teste. Essa informação é de extrema importância no tocante à escolha do melhor método que deve ser, por exemplo, ensinado em cursos de Estatística Básica, ministrados em nível de graduação.

2 Metodologia

Foram comparados, quanto a qualidade do ajuste a densidade normal, os cinco critérios de determinação dos histogramas amostrais citados acima: empírico, de Sturges (1926), de Sturges (1926) modificado por Doane (1976), de Scott (1979) e Freedman e Diaconis (1981). Tal comparação foi realizada por meio de duas etapas distintas. A primeira delas visou determinar as taxas de erro tipo I empíricas do ajuste dos histogramas amostrais de cada critério a distribuição normal. Foram considerados dois níveis nominais de significância (α) usualmente adotados, $\alpha = 1\%$ e $\alpha = 5\%$. Na segunda etapa, a acurácia dos cinco critérios foi avaliada medindo-se o poder em simulações de distribuições sob H_1, ou seja, realizando-se o teste da hipótese nula de normalidade à amostras não-normais geradas sob a hipótese alternativa H_1. Em todas as duas etapas foi utilizado simulação computacional Monte Carlo.

Todas as simulações foram feitas no software estatístico R (R Development Core Team, 2012).

2.1 Algoritmos para obtenção dos histogramas

Para estimar o número de classes (k) dos histogramas amostrais foram utilizados os cinco critérios que foram comparados:

i) Critério de Sturges (1926), em que o número de classes é apresentado em (2),

ii) Critério de Doane (1976), em que o número de classes é apresentado em (3),

iii) Critério empírico, em que o número de classes é apresentado em (4),

iv) Critério de Scott (1979), em que o número de classes é apresentado em (5),

v) Critério de Freedman e Diaconis (1981), em que o número de classes é apresentado em (6).

Após a determinação do número de classes é obtida a amplitude total dos dados:

$$A = X_{(n)} - X_{(1)},$$

em que $X_{(n)}$ é a maior observação e $X_{(1)}$ é a menor observação.

Posteriormente é obtida a amplitude de cada classe (c) por:

$$c = \frac{A}{k-1}.$$

Em seguida, estima-se o limite inferior da primeira classe (LI_1), adotando-se uma correção de "continuidade", dada por:

$$LI_1 = X_{(1)} - \frac{c}{2}.$$

O limite superior da primeira classe é obtido somando a amplitude de classe *(c)* ao limite inferior dessa classe. O limite inferior da classe seguinte é igual ao limite superior da classe anterior. Os limites superiores e inferiores dessa e das demais classes são obtidos desse modo, até completar-se o número de classes.

$$LS_i = LI_i + c \qquad (i = 1, ..., k),$$

$$LI_j = LS_{j-1} \qquad (j = 2, ..., k).$$

Em seguida foram computadas, pela contagem na amostra, as frequências de cada classe, tomando-se por padronização que o limite inferior, nos casos em que ocorreu um valor idêntico na amostra, foi considerado como pertencente a classe e o limite superior não.

2.2 O teste de qui-quadrado

Estipuladas as classes do histograma e suas respectivas frequências, foi testada a seguinte hipótese:

$$\begin{cases} H_0: & \text{Os dados seguem uma distribuição normal;} \\ H_1: & \text{Os dados não seguem uma distribuição normal.} \end{cases} \tag{7}$$

Para isso foram computadas as frequências observadas (FO), admitindo-se que H_0 de (7) é verdadeira. A média e a variância foram estimadas na amostra original e as probabilidades da variável X pertencer a uma determinada classe, foram multiplicadas pelos tamanhos amostrais para se obter as frequências esperadas (FE). Para a primeira classe, no cálculo da frequência esperada o limite inferior foi expandido para $-\infty$ e para a última classe o limite superior foi considerado $+\infty$. O valor da frequência observada da classe i (FO_i) foi comparado com os valores da frequência esperada dessa classe (FE_i) por meio do teste de qui-quadrado dado em (1).

Depois o valor calculado da estatística de qui-quadrado foi comparado com os quantis da distribuição de qui-quadrado, adotando-se nível nominal de significância (α) de 5% e 1%, com $k-1$ graus de liberdade.

2.3 Taxas de erro tipo I

A acurácia dos cinco critérios de obtenção dos histogramas amostrais, subseção 2.1, foi avaliada por meio das taxas de erro tipo I, do ajuste a distribuição normal pelo teste de qui-quadrado, subseção 2.2. Para isso, foi considerada a distribuição normal padrão: $N(\mu = 0, \sigma^2 = 1)$. Foram geradas amostras aleatórias, de diferentes tamanhos amostrais n = 10, 30, 100, 200, 300, 400 e 500. Isto permitiu avaliar o comportamento do teste (taxas de erro tipo I e poder) para pequenos, moderados e grandes tamanhos amostrais.

Geradas as amostras, obtidos os histogramas amostrais e aplicados os testes de qui-quadrado, o processo foi repetido sendo computados os casos em que a hipótese H_0 de (7) foi rejeitada, erro tipo I (rejeitar H_0 sendo ela verdadeira). Foram utilizadas 10.000 simulações de cada tamanho amostral e a proporção de erros tipo I de cada critério computados.

Os procedimentos foram comparados entre si com relação as taxas de significância nominais adotados ($\alpha = 5\%$ e 1%) pelo teste binomial exato bilateral a 5% (LEEMIS e TRIVED, 1996).

2.4 Poder

Para avaliar o poder, foi utilizado um processo semelhante ao anterior, quanto ao número de repetições das simulações, tamanhos amostrais e teste de qui-quadrado, usando H_0 de (7) como hipótese nula. No entanto, para fazer a contagem das frequências observadas (FO) foram consideradas amostras de populações que possuem distribuição exponencial (5), log-normal(0,1) e t de student central com 30 graus de liberdade (JOHNSON e KOTZ, 1970a,b). A proporção de análises em que o teste de qui-quadrado rejeitou H_0, que é falsa, para os níveis nominais de significância, foi considerada uma medida empírica do poder do teste.

2.5 Avaliação da qualidade dos testes

Os critérios que apresentaram taxas de erro tipo I mais próximos dos níveis nominais de significância adotados (5% e 1%) e que apresentaram os maiores valores de poder, para os diferentes tamanhos amostrais e para as diferentes distribuições populacionais, foram considerados ideais.

3 Resultados e discussão

3.1 Taxas de erro tipo I

As Tabelas 1 e 2 contém os resultados de taxa de erro tipo I, cometidos pelos critérios de Doane, Sturges, empírico, Freedman e Diaconis e Scott.

Era esperado que os critérios fossem capazes de desempenhar comportamento assintótico a medida que n aumentasse, mas esse comportamento não ocorreu, ou seja, a taxa de erro tipo I não tende ao valor nominal a medida que se aumenta o tamanho amostral. O valor do erro tipo I começa muito baixo para o tamanho

amostral $n = 10$ e depois vai aumentando passando pelo valor nominal para n entre 30 e 100, a partir daí tende a subir. Esse comportamento ocorre devido a frequência esperada nas classes dos extremos ser baixa, inflacionando assim o valor calculado da estatística qui-quadrado. Dessa maneira o teste fica conservativo para pequenos tamanhos amostrais e conforme se aumenta o tamanho da amostra fica liberal pois o número de classes vai aumentando e o seu comprimento fica menor, diminuindo assim o valor obtido para a frequência esperada.

Tabela 1 - Taxas de erro tipo I cometidas com nível nominal de significância simulado de 1%, para os tamanhos amostrais estudados, para amostras obtidas da distribuição normal padrão $N(\mu = 0, \sigma^2 = 1)$. Os valores em negrito são considerados estatisticamente iguais ao valor nominal 1% pelo teste binomial

Critérios	Tamanhos Amostrais						
	10	30	100	200	300	400	500
Doane	0,0009	0,0139	0,0246	0,0291	0,0299	0,0363	0,0347
Sturges	0,0015	**0,0087**	0,0221	0,0283	0,0357	0,036	0,0342
Crit. Empírico	0,0017	**0,01**	0,0294	0,0393	0,0409	0,0434	0,0467
Freedman e D.	0,0017	**0,0121**	0,032	0,0429	0,0466	0,0482	0,0505
Scott	0,0002	**0,0104**	0,0296	0,0381	0,0426	0,0459	0,0535

Os critérios que figuram com a taxa de erro tipo I mais próximo do valor nominal simulado de $\alpha = 1\%$ são Sturges e Doane, eles apresentam comportamento semelhante ao longo dos tamanhos amostrais com uma ligeira superioridade de Sturges. Os critérios empírico, Diaconis e Freedmann, e Scott apresentaram os piores desempenhos cometendo altas taxas de erro tipo I.

Tabela 2 - Taxas de erro tipo I cometidas com nível nominal de significância simulado de 5%, para os tamanhos amostrais estudados, para amostras obtidas da distribuição normal padrão $N(\mu = 0, \sigma^2 = 1)$. Os valores em negrito são considerados estatisticamente iguais ao valor nominal 5% pelo teste binomial

Critérios	Tamanhos Amostrais						
	10	30	100	200	300	400	500
Doane	0,0074	0,0344	**0,0473**	0,0601	0,0655	0,065	0,0685
Sturges	0,0064	0,0267	**0,0481**	0,0583	0,0635	0,0586	0,063
Crit. Empírico	0,0009	0,0272	0,0586	0,0711	0,075	0,0779	0,0791
Freedman e D.	0,005	0,0315	0,0635	0,0777	0,0878	0,0888	0,0925
Scott	0,0022	0,0263	0,0556	0,073	0,0833	0,0789	0,0828

O comportamento dos critérios para o nível nominal simulado de $\alpha = 5\%$ foi muito semelhante ao anterior, nesta situação novamente os que apresentaram taxas

de erro tipo I mais próximas do valor nominal simulado foram Sturges e Doane. Diaconis e Freedmann chegou a cometer 9% de erro para $n = 500$.

O critério de determinação do valor de k que melhor controlou a taxa de erro tipo I foi curiosamente o critério mais antigo dentre os estudados (Sturges, 1926). Este critério é o que apresenta esta taxa mais próxima dos valores nominais simulados para tamanhos amostrais mais "normais", sendo esta taxa igual ao valor nominal para $n = 30$ em $\alpha = 1\%$ e para $n = 100$ em $\alpha = 5\%$, segundo o teste binomial exato bilateral a 5%.

3.2 Poder

O poder dos critérios de construção de histogramas foi medido sobre a hipótese alternativa de não-normalidade. Para isso, foram simuladas amostras das seguintes distribuições com seu respectivos valores paramétricos: exponencial (5), log-normal (0,1) e t de student (30). Foram utilizados todos os tamanhos amostrais já mencionados e para cada situação, foi aplicado o teste qui-quadrado com nível nominal de significância de 5% e 1%.

Na tabela 3 são apresentados os níveis de poder atingidos pelos cinco critérios de obtenção do número de classes, para todas as combinações das três distribuições, de todos os tamanhos amostrais e do teste qui-quadrado a 5% e a 1%.

As duas primeiras situações se referem ao ajuste da distribuição exponencial (5) que, por ser muito distinta da distribuição normal, apresenta altas taxas de poder. A taxa de poder começa baixa para a menor amostra simulada ($n = 10$) mas cresce rapidamente e para $n \geq 200$ todos os critérios já apresentam 100% de poder.

Nas próximas situações que se referem a distribuição log-normal(0,1), os níveis de poder são maiores que os cometidos sob a distribuição exponencial, fato que ocorre devido a grande assimetria desta distribuição. Os valores novamente começam baixos para o tamanho amostral simulado $n = 10$ mas crescem rapidamente, muito mais rápido do que na distribuição exponencial, e chegam a atingir o 100% de poder para $n \geq 200$. Mas pode ser observado também que para $n = 100$ alguns critérios já figuram extremamente próximos de alcançar os 100% de poder.

Nos dois últimos casos estão as taxas de poder cometidas pelo teste qui-quadrado no ajuste da distribuição t de Student (30), distribuição muito semelhante a distribuição normal, sendo igual à normal com infinitos graus de liberdade. Sendo que o número de graus de liberdade 30 é considerado o limite, a partir daí a distribuição normal e a distribuição t de Student já se confundem.

Já eram esperadas na distribuição t-student(30) taxas de poder mais baixas que nos ajustes anteriores, justamente pela facilidade de confundí-la com a distribuição normal pelas regras de construção de histogramas. Nesta distribuição melhores comparações podem ser feitas, pois o padrão de crescimento das taxas de poder ao longo dos tamanhos amostrais é diferente para cada critério.

Tabela 3 - Taxas de poder cometidas pelos critérios de determinação de k ao longo de todos os tamanhos amostrais para distribuições Exponencial(5), Log-normal(0,1) e t-student(30) para o teste qui-quadrado com nível nominal de significância de 5% e 1%

	10	30	100	200	300	400	500
Exponencial 5%							
Doane	0,0769	0,5993	0,9908	1	1	1	1
Sturges	0,0689	0,5584	0,9916	1	1	1	1
Crit. Empírico	0,0398	0,5689	0,9954	1	1	1	1
Freedman e Diaconis	0,0856	0,582	0,9908	1	1	1	1
Scott	0,0373	0,5436	0,9886	1	1	1	1
Exponencial 1%							
Doane	0,0262	0,4746	0,9791	1	1	1	1
Sturges	0,0087	0,4492	0,9712	1	1	1	1
Crit. Empírico	0,0046	0,4389	0,9824	1	1	1	1
Freedman e Diaconis	0,335	0,4531	0,9734	1	1	1	1
Scott	0,004	0,4104	0,9691	1	1	1	1
Log-normal 5%							
Doane	0,187	0,8489	0,9999	1	1	1	1
Sturges	0,171	0,8268	0,9997	1	1	1	1
Crit. Empírico	0,1181	0,8237	0,9999	1	1	1	1
Freedman e Diaconis	0,2011	0,8335	0,9999	1	1	1	1
Scott	0,1167	0,8109	0,9996	1	1	1	1
Log-normal 1%							
Doane	0,0933	0,7621	0,9996	1	1	1	1
Sturges	0,05	0,7357	0,9994	1	1	1	1
Crit. Empírico	0,0302	0,739	0,9999	1	1	1	1
Freedman e Diaconis	0,1169	0,7545	0,9996	1	1	1	1
Scott	0,0259	0,717	0,9995	1	1	1	1
t-student 5%							
Doane	0,0076	0,044	0,0944	0,1369	0,1696	0,1787	0,2095
Sturges	0,0084	0,0387	0,1028	0,1322	0,1692	0,1897	0,1964
Crit. Empírico	0,007	0,0418	0,1116	0,1557	0,1822	0,2128	0,2391
Freedman e Diaconis	0,0058	0,0428	0,1193	0,1648	0,2018	0,2306	0,2397
Scott	0,0025	0,0373	0,1055	0,1653	0,1934	0,2254	0,2396
t-student 1%							
Doane	0,0016	0,0262	0,0603	0,0879	0,113	0,1297	0,1494
Sturges	0,0104	0,0186	0,0595	0,0944	0,1131	0,1342	0,1352
Crit. Empírico	0,0104	0,0155	0,0731	0,1101	0,1333	0,1492	0,1639
Freedman e Diaconis	0,0039	0,0271	0,0764	0,1206	0,1499	0,1731	0,1881
Scott	0,0102	0,0154	0,0681	0,1089	0,1424	0,1538	0,1698

A maior taxa de poder observada nesta distribuição para o teste qui-quadrado realizado ao nível nominal de significância de 5% foi apresentada pelo critério de Freedman e Diaconis chegando a atingir $23,97\%$ para $n = 500$. Esse critério já apresenta superioridade desde as pequenas amostras, comportamento que é

explicado pelo fato da taxa de erro tipo I ser diretamente proporcional ao poder e como pode ser observado nas Tabelas 1 e 2 Freedman e Diaconis apresentaram altas taxas de erro tipo I implicando em altos níveis de poder.

Os critérios de Scott e empírico apresentam comportamento semelhante, sendo que o critério empírico figura com melhores resultados para tamanhos amostrais pequenos ($n < 200$) e Scott foi melhor para tamanhos amostrais maiores. Sturges e Doane ficaram com os piores desempenhos nessa situação, chegando a atingir no máximo 19,64% e 20,95% de poder, respectivamente. Conforme pode ser observado nas tabelas 1 e 2, Sturges apresenta as mais baixas taxas de erro tipo I, fato que ocasiona este baixo poder.

Quando o nível nominal de significância adotado foi de 1%, os resultados para a distribuição t-student também foram semelhantes, com o critério de Freedman e Diaconis atingindo os melhores níveis, e Doane e Sturges ficando com os piores desempenhos.

Para amostras geradas de uma distribuição exponencial(5), que é muito diferente da distribuição normal, a taxa de poder atinge o máximo em todos os critérios para $n \geq 200$ e o comportamento deles para tamanhos amostrais menores acaba sendo muito parecidos impossibilitando assim uma melhor comparação. Situação semelhante ocorre quando a população em questão é a log-normal(0,1), onde a taxa de poder também atinge o máximo para amostras grandes "$n \geq 200$".

Nas amostras da distribuição t-student(30), que foi escolhida de propósito por ser muito parecida com a normal, torna-se possível uma melhor comparação entre as taxas de poder cometidas pelos diferentes critérios. O critério de Freedman e Diaconis (1981) é o que apresenta os melhores resultados, tanto para o nível nominal de significância de 5% quanto para 1%.

Conclusões

Os resultados do teste de qui-quadrado são influenciados pela escolha do critério de determinação do número de classes, que é o primeiro passo na execução deste teste, pois os critérios estudados apresentaram comportamentos diferentes ao longo dos tamanhos amostrais simulados.

Com base nos resultados obtidos é aconselhável utilizar, quando o objetivo for controlar o erro tipo I por meio do teste de qui-quadrado, o valor de k determinado pela regra de Sturges (1926), pois com este critério o teste controla melhor as taxas de erro tipo I. Se o objetivo for o poder, aconselha-se utilizar k determinado pelo critério de Freedman e Diaconis (1981), pois com este critério o teste apresenta as taxas de poder mais altas dentre os critérios estudados.

Esta divisão dos resultados, ora utilizando um critério, ora outro conforme se muda o objetivo (erro tipo I ou poder), já era esperada, pois conseguir um critério que apresente bons resultados tanto para poder quanto para erro tipo I é difícil uma vez que estas taxas são diretamente proporcionais. O que usualmente é feito para evitar esta divisão na recomendação, é escolher um critério que controle a taxa de erro tipo I e caso esteja ineteressado em maiores valores do poder,

aumentar a significância nominal adotada (por exemplo, de $\alpha = 5\%$ para $\alpha = 10\%$), aumentando assim seu poder. Dessa forma, caso deva ser recomendado apenas um critério, aconselha-se utilizar o critério de Sturges (1926), pois este critério é o que melhor controla as taxas de erro tipo I.

Agradecimentos

Agradecimentos ao CNPq, pela bolsa de mestrado concedida ao primeiro autor.

FERNANDES, T. J.; FERREIRA, E. B.; FERREIRA, D. F. Evaluation Monte Carlo of the test of chi-square normality under different criteria of the number of classes. *Rev. Bras. Biom.*, São Paulo, v.30, n.2, p.185-198, 2012.

▪ *ABSTRACT: It is known that the probability density function can be approximated by the sample histograms. The normal density is the most used due to such assumption in most statistical tests. One way to verify the normality of the data is through the adherence chi-square test. However, authors usually do not explain what was the rule used to determine the number of classes of the histogram (which is the first step in implementing this test). Thus, a possible poor performance is attributed to the chi-square test, but one suspects that it may be increased by choosing a non optimal criterion for determining the number of classes (k). The objectives of this study were: 1. verify the validity of this hypothesis and; 2. point out what is the better criterion. We compared the performance (type I error rate and power) of five criteria for the determination of k (empirical criteria, Sturges, Scott, Doane and Freedman and Diaconis) for the normality chi-sqaure test, via Monte Carlo simulation. It was concluded that the results of the chi-square test are affected by the criterion for k and that the Sturges criterion best controlled the type I error rate. The criterion of Freedman and Diaconis presented the highest values of power.*

▪ *KEYWORDS: Adherence test; type I error rate; power; histogram.*

Referências

BROWN, L. D.; HWANG, J.T. How to aproximate a histogram by a normal density. *Am. Stat.*, Baltimore, v.47, n.4, p.251-255, 1993.

CABRAL, A. P. *Influência de fatores ambientais na Leishmaniose visceral no Rio Grande do Norte*. 2007. Dissertação (Mestrado) – Universidade Federal do Rio Grande do Norte, Natal, 2007.

CAMPOS, H. *Estatística experimental não-paramétrica.* 4.ed. Piracicaba: FEALQ, 1983. 349p.

CAMPOS, T. F.; RODRIGUES, C. A.; FARIAS, I. M. A.; RIBEIRO, T. S.; MELO, L. P. Comparação dos instrumentos de avaliação do sono, cognição e função no acidente vascular encefálico com a classificação internacional de funcionalidade, incapacidade e saúde (CIF), *Rev. Bras. Fisioter.*, São Carlos, v.16, n.1, p.23-29, 2012.

CIELASK, F.; LEVANDOSKI, G.; GÓES, S. M.; SANTOS, T. K.; VILELA JÚNIOR, G. B.; LEITE, N. Relação do nível de qualidade de vida e atividade física em acadêmicos de educação física. *Fit. Perf. J.*, Rio de Janeiro, v.6, n.6, p.357-361, 2007.

CIRILO, M. A.; FERREIRA, D. F. Extensão do teste para normalidade univariado baseado no coeficiente de correlação quantil-quantil para o caso multivariado. *Rev. Mat. Estat.*, São Paulo, v.21, n.3, p.67-84, 2003.

DOANE, D. P. Aesthetic frequency classifications. *Am. Stat.*, Baltimore, v.30, n.4, p.181-183, 1976.

FERREIRA, D. F. *Estatística básica*. Lavras: UFLA, 2005. 664p.

FREEDMAN, D; DIACONIS, P. On the histogram as a density estimator: L2 theory. *J. Probab. Theory Relat. Fields.*, New York, v.57, n.4. p.453-476. 1981.

JOHNSON, N. L.; KOTZ, S. *Continuous univariate distributions-1.* New York: John Wiley & Sons, 1970a. 300p.

JOHNSON, N. L.; KOTZ, S. *Continuous univariate distributions-2.* New York: John Wiley & Sons, 1970b. 306p.

LEEMIS, L. M.; TRIVEDI, K. S. A comparison of approximate interval estimators for the Bernoulli parameter. *Am. Stat.*, Baltimore, v.50, n.1, p.63-68, 1996.

LINHARES, R. S.; HORTA, B. L.; GIGANTE, D. P.; COSTA, J. S. D.; OLINTO, M. T. A. Distribuição de obesidade geral e abdominal em adultos de uma cidade no Sul do Brasil. *Cad. Saúde Pública*, Rio de Janeiro, v.28, n.3, p.438-448, 2012.

LOPES, H. E. G. Abrindo a caixa preta: considerações sobre a utilização da Análise Fatorial Confirmatória nas pesquisas em administração. *Rev. Econ. Gest.*, Belo Horizonte, v.5, n.11, p.19-34, 2005.

LUNZ, W.; MOLINA, M. D. C. B.; RODRIGUES, S. L.; GONÇALVES, C. P.; BALDO, M. P.; VIANA, E. C.; DANTA, E. M.; MILL, J. G. Impacto da atividade física sobre o risco cardiovascular na população adulta de Vitória - ES. *Rev. Bras. Ciên. Mov.*, Brasília, v.18, n.3, p.64-73, 2010.

MCCULLAGH, P., NELDER, J. A. *Generalized linear models.* 2.ed. Cambridge: Chapman & Hall, 1989. 511p.

MENDES, M.; PALA, A. Type I error rate and power of three normality tests. *Pak. J. Inf. Technol.*, Faisalabad, v.2, n.2, p.135-139, 2003.

MOTA, T. S; SILVEIRA, L. V. A.; ANTUNES, A. A. Modelo de Cox para eventos cardiovasculares recorrentes em pacisntes sob diálise com covariáveis medidas no tempo. *Rev. Bras. Biom.*, São Paulo, v.30, n.1, p.150-159, 2012.

MORAES, C. C. G.; GUERREIRO, A. N.; KURODA, R. B. S.; MENESES, A. M. C.; VASCONCELLOS, S. A. Inquérito sorológico para leptospirose em rebanhos de ovinos no município de Igarapé-Açu, Estado do Pará. *Rev. Ciên. Agrár.*, Lisboa, v.55, n.1, p.58-60, 2012.

NEVES, A. C. C.; MONTEIRO, A. M.; NG, H. G. Prevalência das fissuras labiopalatinas na Associação de Fissurados Labiopalatinos de São José dos Campos/SP. *Rev. Bioc.*, Taubaté, v.8, n.2, p.69-74, 2002.

PEDROSO, J. A. R.; PASKULIN, D. A.; DIAS, F. S.; FRANÇA, E.; ALHO, C. S. Análise da tendência temporal de dano renal agudo entre pacientes graves conforme polimorfismos I/D e-262A > T da enzima conversora da angiotensina *J. Bras. Nefrol.*, São Paulo, v.32, n.2, p.182-194, 2010.

PINHAT E. C.; BORBA, M. G.; FERREIRA, M. L.; FERREIRA M. A.; FERNANDES, R. K.; NICOLAOU, S. K.; OKAMOTO, C. T.; NETO, C. F. O. Fungal colonization in newborn babies of very low birth weight: a cohort study. *J. Pediat.*, Rio de Janeiro, v.88, n.3, p.211-216, 2012.

R DEVELOPMENT CORE TEAM. *R*: a language and environment for statistical computing. R Foundation for Statistical Computing. Viena, Austria. Disponível em: <http://www.R-project.org>. Acesso em: 10 fev. 2012.

ROSASCO, F. V. Estudo sazonal e caracterização do aerossol black carbon medido no Inpe de São José dos Campos. São José dos Campos: Inpe, 2009. p.157.

SARCINELLI, T. S.; FERNANDES FILHO, E. I.; SCHAEFER, C. E. G. R.; DEMARCO JÚNIOR, P.; LEITE, F. P. Representatividade Fisiológica e Pedológica de fragmentos de floresta nativa em áreas de plantios homgêneos de eucalipto, *Rev. Árvore*, Viçosa, v.36, n.3, p.499-509, 2012.

SCOTT, D. W. On optimal and data-based histograms. *Biometrika*, Oxford, v.66, n.3, p.605-610, 1979.

SEIER, E. *Comparison of tests for univariate normality.* Johnson City: Department of Mathematics, East Tennessee State University, 2004. 17p.

SILVA, C. T.; FERREIRA, E. B. Desempenho de testes de normalidade via simulação Monte Carlo. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETIA – RBRAS, 54., 2009, São Carlos. *Anais...* p.1-5.

STURGES, H. A. The choice of a class interval. *J. Am. Stat. Assoc.*, 65p. 1926.

TORMAN, V. B. L.; COSTER, R.; RIBOLDI, J. Normalidade de variáveis: métodos de verifcação e comparação de alguns testes não paramétricos por simulação. , *Rev. HCPA*, Porto Alegre, v.32, n.2, p.227-234, 2012.

YEH, W. S. C.; KIONARA, P. R.; SOARES, I. S. C.; CARMONA, M. J. C.; ROCHA, F. T.; GALVÃO, C. E. S. Prevalência de sinais de sensibilidade ao látex em pacientes com mielomeningocele submetidos a múltiplos procedimentos cirúrgicos. *Rev. Bras. Anestesiol.*, v.62, n.1, p.56-62, 2012.

Recebido em 11.04.2012.

Aprovado após revisão em 08.08.2012.