TESTE MONTE CARLO DE NORMALIDADE UNIVARIADO BASEADO EM DISTÂNCIAS

Nelson de Almeida PEREIRA FILHO[1]
Daniel Furtado FERREIRA[2]

▪ RESUMO: As distribuições normais de probabilidade descrevem o comportamento de muitos fenômenos da vida real em vários campos da ciência. Ao se retirar uma amostra aleatória de uma população, no caso univariado, é comum se pressupor que os dados ou resíduos do modelo adotado são proveniente de uma população normalmente distribuída. Os gráficos, como histogramas e gráficos Q-Q, são maneiras bastante eficientes, porém subjetivas de se verificar a normalidade da distribuição dos dados ou dos resíduos do modelo considerado. No entanto, isso não é suficiente para se fazer inferência sobre a normalidade dos dados coletados ou dos resíduos de modelos ajustados. Existem inúmeros testes de normalidade na literatura. Entre eles o teste de Shapiro-Wilk, considerado como tendo propriedades ótimas. Entretanto, esse teste possui a limitação computacional de ser aplicável a um número de observações inferiores a 5.000. Este artigo tem como objetivo propor um teste de normalidade univariada, baseado nas distâncias entre os valores esperados das estatísticas de ordem dos valores observados na amostra e os valores esperados das estatísticas de ordem da distribuição normal padrão, que possa ser usado em quaisquer tamanhos de amostras. Também objetivou comparar o teste de normalidade Shapiro-Wilk com o teste de normalidade univariado proposto. A distribuição da estatística foi obtida via simulação Monte Carlo. Os resultados obtidos de poder e controle do erro tipo I, permitem que se conclua que a proposta é, em geral, mais eficiente que o teste Shapiro-Wilk e não possui a limitação prática de ser restrito a tamanho de amostra máximo de 5.000 unidades.

▪ PALAVRAS-CHAVE: Normalidade; distâncias; poder; Shapiro-Wilk.

[1]Instituto Federal da Bahia – IFBA, Departamento de Ciências Aplicadas, CEP: 40110 -150, Salvador, BA, Brasil. E-mail: *nelson@ifba.edu.br*
[2]Universidade Federal de Lavras – UFLA, Departamento de Ciências Exatas, Caixa Postal 37, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *danielff@dex.ufla.br*, Bolsista CNPq.

1 Introdução

A estatística inferencial é um ramo da estatística que consiste em fazer afirmativas válidas sobre parâmetros de alguma população, baseadas em dados amostrais. Ao se fazer uma estimativa pontual acerca de um certo parâmetro populacional, sabe-se que ela tem grande probabilidade de não ser igual ao parâmetro populacional. Deste modo, o processo de inferência por meio de estimação intervalar e testes de hipóteses deve ser usado. Ao se retirar uma amostra de uma população, em grande parte dos casos, pressupõe-se que os dados sejam provenientes de uma população normalmente distribuída. Tal suposição é feita pelo simples fato de que as distribuições normais podem ser usadas para descrever muitas situações da vida real e são largamente aplicadas em vários campos da ciência. Além disso, a facilidade da obtenção de intervalos de confiança e testes exatos é grande, quando se assume esse modelo para a distribuição dos dados ou dos resíduos do modelo subjacente. Ademais, a maior parte dos testes e procedimentos de estimação existentes é formulada tomando-se o modelo normal com referência para a distribuição dos dados.

Deste modo, ao se retirar uma amostra aleatória de uma população, no caso univariado, assumimos geralmente que os dados são provenientes de uma população normalmente distribuída. Uma maneira simples, porém subjetiva, de se verificar a normalidade de uma distribuição é por meio de gráficos, como histogramas e gráficos Q-Q (Filliben, 1975). Tal observação, no entanto, não é suficiente pra se fazer inferência sobre a normalidade. No caso multivariado, principalmente nas situações de muitas variáveis, isso se torna ainda mais complicado porque nem sempre é possível detectar-se alguma violação da normalidade, haja vista a complexa relação existentes entre as varáveis envolvidas. Apesar disso, os gráficos Q-Q, são ferramentas viáveis para a visualização de valores discrepantes da amostra. Inúmeros testes de normalidade univariada existem. Entretanto, apesar de a maioria controlar adequadamente os erros tipo I, o poder desses testes variam consideravelmente em função das diferentes distribuições não-normais consideradas em suas avaliações (Oliveira e Ferreira, 2010).

A ausência de um teste uniformemente poderoso em relação aos tamanhos amostrais e distribuições tem sido o principal motivo da grande quantidade de proposições de testes de normalidade.

Um teste que vem sendo largamente utilizado como referência, tanto nas aplicações científicas em geral quanto nos trabalhos de comparação de desempenho de outros testes propostos é o de Shapiro-Wilk. A estatística do teste de Shapiro-Wilk é denotada por W, sendo calculada por

$$W = \frac{\left[\sum_{i=1}^{n} a_i X_{(i)}\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

em que as constantes a_1,a_2, ...,a_n, elementos do vetor $\boldsymbol{a}$, são calculadas como solução de

$$\boldsymbol{a} = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}}{(\boldsymbol{m}^\top \boldsymbol{V}^{-2} \boldsymbol{m})^{\frac{1}{2}}}, \tag{1}$$

em que $X_{(i)}$ é a i-ésima estatística de ordem amostral, $\bar{X}$ é a média amostral, $\boldsymbol{m}$ é o vetor $n \times 1$ dos valores esperados da estatísticas de ordem da normal padrão e $\boldsymbol{V}$, a matriz $n \times n$ das variâncias e covariâncias das estatísticas de ordem da normal padrão.

O valor esperado da i-ésima estatística de ordem da normal padrão m_i, i-ésimo componente de $\boldsymbol{m}$, é muito bem estimado por

$$\tilde{m}_i = \Phi^{-1}\left(\frac{j - 3/8}{n + 1/4}\right), \tag{2}$$

em que $\Phi^{-1}(p)$ é a inversa da função de distribuição da normal padrão avaliada no argumento p, entre 0 e 1. O argumento $(j - 3/8)/(n + 1/4)$ de (2), $j = 1, 2, \ldots, n$, representa a função de distribuição empírica com correções de continuidade para se obter uma melhor aproximação, sendo dadas por 3/8, no numerador, e 1/4, no denominador. O vetor $\boldsymbol{m}$ é, portanto, estimado pelo vetor $(n \times 1)$ $\tilde{\boldsymbol{m}} = [\tilde{m}_i]$.

O vetor de coeficientes $\boldsymbol{a}$ pode ser calculado utilizando várias aproximações, desde as de Shapiro e Wilk (1965), até as de Royston (1982, 1992, 1993, 1995). As aproximações apresentadas por Shapiro e Wilk (1965) e Royston (1982) não são precisas. As aproximações de Royston (1992, 1993) para o vetor de coeficientes $\boldsymbol{a}$ devem ser utilizadas para $n \geq 4$. Esta aproximação é baseada nas médias das estatísticas de ordem que são estimadas pela equação (2). Para $n = 3$, o vetor de coeficientes $\boldsymbol{a}$ é obtido de forma exata.

Para se estimar o vetor de coeficientes $\boldsymbol{a}$, Royston (1993) apresenta o seguinte roteiro. Deve-se inicialmente obter

$$\begin{aligned} \tilde{a}_n = & c_n + 0{,}221157u - 0{,}14798u^2 - 2{,}071190u^3 \\ & + 4{,}434685u^4 - 2{,}706056u^5 \end{aligned} \tag{3}$$

$$\begin{aligned} \tilde{a}_{n-1} = & c_{n-1} + 0{,}042981u - 0{,}293762u^2 - 1{,}752461u^3 \\ & + 5{,}682633u^4 - 3{,}582633u^5, \end{aligned}$$

em que c_n e c_{n-1} são o n-ésimo e $(n-1)$-ésimo elementos do vetor $\boldsymbol{c} = (\tilde{\boldsymbol{m}}^\top \tilde{\boldsymbol{m}})^{-1/2} \tilde{\boldsymbol{m}}$ e $u = 1/\sqrt{n}$.

Deve-se obter a quantidade normalizadora

$$\phi = \begin{cases} (\tilde{\boldsymbol{m}}^\top \tilde{\boldsymbol{m}} - 2\tilde{m}_n^2)/(1 - 2\tilde{a}_n^2) & \text{se } n \leq 5 \\ (\tilde{\boldsymbol{m}}^\top \tilde{\boldsymbol{m}} - 2\tilde{m}_n^2 - 2\tilde{m}_{n-1}^2)/(1 - 2\tilde{a}_n^2 - 2\tilde{a}_{n-1}^2) & \text{se } n > 5 \end{cases}$$

e finalmente

$$\tilde{a}_j = \frac{\tilde{m}_j}{\sqrt{\phi}}, \tag{4}$$

para $j = 2, \ldots, n-1$ ($n \leq 5$) ou $j = 3, 4, \ldots, n-2$ ($n > 5$). Deve-se observar que $\tilde{a}_1 = -\tilde{a}_n$ e $\tilde{a}_2 = -\tilde{a}_{n-1}$.

A estatística do teste Shapiro-Wilk é redefinida a partir das estatísticas de ordem $X_{(j)}$ e do vetor de coeficientes $\tilde{\boldsymbol{a}}$ ($n \times 1$) da seguinte forma

$$W = \frac{\left(\sum_{j=1}^{n} \tilde{a}_j X_{(j)}\right)^2}{\sum_{j=1}^{n} \left(X_j - \bar{X}_\cdot\right)^2}. \tag{5}$$

A estatística W não segue uma distribuição normal, mas Royston (1993) propõe a utilização de uma transformação da família Box-Cox, para obter normalidade. De acordo com a proposta de Royston (1993), deve-se obter Y (valor transformado de W) por

$$Y = \begin{cases} -\ln\left[\gamma - \ln(1-W)\right] & \text{se } 4 \leq n \leq 11 \\ \ln(1-W) & \text{se } 12 \leq n \leq 5.000 \end{cases} \tag{6}$$

sendo $\gamma = -2{,}273 + 0{,}459n$.

A variável Y possui média dada por

$$\mu_Y = \begin{cases} 0{,}5440 - 0{,}39978n + 0{,}025054n^2 - 0{,}0006714n^3 & \text{se } 4 \leq n \leq 11 \\ -1{,}5861 - 0{,}31082u - 0{,}083751u^2 + 0{,}0038915u^3 & \text{se } 12 \leq n \leq 5.000 \end{cases} \tag{7}$$

em que $u = \ln(n)$ e desvio padrão

$$\sigma_Y = \begin{cases} \exp\left\{1{,}3822 - 0{,}77857n + 0{,}062767n^2 - 0{,}0020322n^3\right\} & \text{se } 4 \leq n \leq 11 \\ \exp\left\{-0{,}4803 - 0{,}082676u + 0{,}0030302u^2\right\} & \text{se } 12 \leq n \leq 5.000 \end{cases} \tag{8}$$

Assim, sob a hipótese nula de normalidade

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \tag{9}$$

possui aproximadamente distribuição normal-padrão e o valor-p é estimado por valor-$p = 1 - \Phi(Z)$, ou seja, o valor-p corresponde a área da distribuição normal-padrão à direita de Z.

Para o caso particular de $n = 3$, o vetor $\boldsymbol{a}$ é conhecido e exato, sendo dado por $\boldsymbol{a} = [-\sqrt{2}/2,\ 0,\ \sqrt{2}/2]^\top$. O valor-$p$ associado a W é calculado de forma exata por

$$\text{valor-}p = 1 - F(w) = 1 - \frac{6}{\pi}\left[\text{arcsen}(\sqrt{w}) - \text{arcsen}\left(\sqrt{\frac{3}{4}}\right)\right]$$

em que a função arcsen deve ser tomada em radianos.

Por meio da comparação do valor-p com o nível nominal de significância α adotado, toma-se a decisão de rejeitar ou não a hipótese nula de normalidade. Convém salientar, como pode ser observado nas aproximações de Royston (1993), que só se deve aplicar o teste para $n \leq 5.000$. Isso se deve ao fato de que as aproximações obtidas por este autor foram validadas apenas para esse limite máximo do tamanho amostral.

A validade da inferência, para a maioria dos procedimentos que supõe normalidade é garantida quando essa distribuição é de fato a distribuição dos dados ou dos resíduos de um modelo linear. Um limitação ao uso desse teste é a limitação dos tamanhos amostrais ao máximo de 5.000 observações. Deve-se salientar que essa limitação é de natureza computacional e não teórica, uma vez que Royston (1993) não validou as suas aproximações para amostras maiores que esse limite, talvez pelas restrições de *hardware* existentes na ocasião da realização do estudo. Em muitas aplicações em zootecnia, na área de melhoramento animal, e em ciências florestais, é muito comum o pesquisador se deparar com amostras superiores a esse tamanho. Existem alternativas de testes que são aplicáveis às grandes amostras, mas possuem limitações. Dentre esses testes pode-se citar o teste de Kolmogorov-Smirnov e os testes baseados em assimetria e curtose. O teste de Kolmogorov-Smirnov é apontado por Thode (2002) como tendo baixo poder. Já os testes de assimetria e curtose possuem limitações referentes à não-garantia de que exista normalidade pelo simples fato de não haver desvios de simetria ou de mesocurtose, respectivamente (Thode, 2002). Além disso, a distribuição de suas estatísticas são válidas apenas assintoticamente. Sendo assim, o uso de testes estatísticos se fazem necessários para se inferir sobre a normalidade da distribuição dos dados ou resíduos que estão sendo analisados, que possam ser aplicados em grandes amostras.

Recentemente, Yazici e Yolocan (2007) realizaram uma comparação entre 15 testes de normalidade usando simulações de Monte Carlo. O resultado da comparação foi, segundo eles, que o poder, a facilidade de uso e consequentemente a escolha do teste, depende de vários fatores, entre eles o tipo de distribuição sob H_1, o tamanho da amostra e os valores críticos. Embora os autores apontem algumas limitações do teste de Shapiro-Wilk dadas pela necessidade de estimar coeficientes e valores críticos especiais, eles afirmam que este teste fornece um elevado poder para detectar a não-normalidade sob a hipótese alternativa considerando várias distribuições simétricas, não-simétricas, caudas pesada ou leves e sobre todos os tamanhos de amostras utilizados. Um outro trabalho, apresentado por Romão, Delgado e Costa (2010), compara 33 testes de normalidade e não aponta um único teste com sendo mais poderoso do que os demais em todos os casos estudados. Eles

classificam os testes de acordo com características comuns, em geral grupos de três ou quatro testes e, em cada grupo, apontam o teste de maior poder. Especificamente entre os testes mais poderosos para distribuições assimétricas e distribuições que são misturas de normais ou normais com presença de *outliers*, os autores recomendam o teste W de Shapiro-Wilk. Quando a natureza da não-normalidade não é conhecida *a priori*, o teste de Shapiro-Wilk está entre os três testes recomendado pelos autores.

O objetivo do presente trabalho é propor um teste baseado nas distâncias entre os valores observados das estatísticas de ordem amostrais e os valores esperados dessas mesmas estatísticas de ordem na distribuição normal padrão e que possa ser aplicado para amostras maiores que 5.000. Também objetivou-se comparar o desempenho do teste de normalidade Shapiro-Wilk com o do teste de normalidade univariado proposto, por meio do uso de simulação Monte Carlo.

2 Metodologia

No presente trabalho é proposto um teste computacionalmente intensivo baseado em simulação Monte Carlo para a hipótese nula

$$H_0 : f(x) = \phi_0(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{10}$$

contra a alternativa

$$H_1 : f(x) \neq \phi_0(x).$$

Deve ficar claro, que a hipótese nula é do tipo composta, ou seja, somente a forma (família) da distribuição é definida em (10), mas os parâmetros são desconhecidos. Assim, nas simulações Monte Carlo, os valores x_1, x_2, ..., x_n são realizações de uma variável aleatória X pertencente à família de funções densidade de probabilidade da normal $f(x)$.

2.1 Proposta

Inicialmente foi considerado que uma amostra aleatória X_1, X_2, ..., X_n, de tamanho n, supostamente obtida sob H_0, é extraída. Em seguida são obtidas as estatísticas de ordem $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$, ou seja, a amostra é ordenada de forma crescente.

O teste proposto adota como estatística a distância entre os valores observados das estatísticas de ordem na amostra e os valores esperados das estatísticas de ordem da distribuição normal padrão. A distribuição dessa estatística do teste foi obtida via simulação Monte Carlo, assumindo que o modelo normal determinado em (10) é o modelo dos dados submetidos ao teste. As distâncias, considerando a amostra original, d_c^2, são distâncias obtidas entre os desvios da j-ésima estatística de ordem $X_{(j)}$ da média amostral $\bar{X}$ normalizadas e o seu valor esperado estimado supondo

normalidade $\tilde{m}_j$, também normalizado. Essa distância, para vetores $\boldsymbol{x}$ e $\boldsymbol{y}$ de mesma dimensão, é calculada por meio da expressão

$$d^2(\boldsymbol{x},\boldsymbol{y}) = \left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\right)^\top \left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\right), \tag{11}$$

em que $\|\boldsymbol{x}\|$ e $\|\boldsymbol{y}\|$ representam as normas euclidianas dos vetores $\boldsymbol{x}$ e $\boldsymbol{y}$, respectivamente, ou seja, $\|\boldsymbol{x}\| = \sqrt{\sum_{j=1}^{n} x_{(j)}^2}$. Para o teste em questão, o vetor $\boldsymbol{x}$ é formado pelos desvios das estatísticas de ordem amostrais em relação a média amostral e o vetor $\boldsymbol{y}$, pelos valores esperados e estimados das estatísticas de ordem normais padrão.

Sendo $\boldsymbol{X} = [X_{(1)}, X_{(2)}, \ldots, X_{(n)}]^\top$, o vetor aleatório n-dimensional contendo as estatísticas de ordem observadas, $\tilde{\boldsymbol{m}} = [\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_n]^\top$, o estimador do valor esperado do vetor das estatísticas de ordem da normal padrão e $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \bar{X}\boldsymbol{1}$, em que $\boldsymbol{1} = [1, 1, \ldots, 1]^\top$ é o vetor unitário e $\bar{X} = \sum_{j=1}^{n} X_j/n$, então a expressão da estatística do teste a ser utilizada no contexto da proposta é

$$d^2(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{m}}) = \left(\frac{\tilde{\boldsymbol{X}}}{\|\tilde{\boldsymbol{X}}\|} - \frac{\tilde{\boldsymbol{m}}}{\|\tilde{\boldsymbol{m}}\|}\right)^\top \left(\frac{\tilde{\boldsymbol{X}}}{\|\tilde{\boldsymbol{X}}\|} - \frac{\tilde{\boldsymbol{m}}}{\|\tilde{\boldsymbol{m}}\|}\right). \tag{12}$$

Os valores esperados $E\left(Z_{(j)}\right) = m_j$, $i = 1, 2, \ldots, n$, possuem cálculo difícil de ser obtido numericamente por envolver distribuições de estatísticas de ordem. Dessa forma, a seguinte aproximação (Royston, 1993) é utilizada

$$\tilde{m}_i = \Phi^{-1}\left(\frac{j - 3/8}{n + 1/4}\right), \tag{13}$$

em que $\Phi^{-1}(\cdot)$ é a inversa da função de distribuição da normal padrão.

O próximo passo é utilizar simulações Monte Carlo, considerando um número grande de repetições N_{MC}, para obtenção da distribuição nula da estatística do teste. Para isso amostras de tamanho n com distribuições normais são geradas, sendo calculadas as distâncias entre os valores das estatísticas de ordem amostrais e os valores esperados das estatísticas de ordem na distribuição normal padrão, estimados utilizando (13). Repetindo o processo N_{MC} vezes e armazenando as distâncias d_c^2's, a distribuição nula é obtida. O teste será concretizado calculando o valor-p, pelo dobro do mínimo das proporções de distâncias da distribuição nula de Monte Carlo que são ou superiores ou inferiores a distância obtida considerando a amostra original. Considerando d_j^2, a j-ésima distância obtida na j-ésima simulação, para $j = 1, 2, \ldots, N_{MC} + 1$, em que a estatística original é incluída, o valor-p é obtido por

$$\text{valor-}p = \frac{\sum_{j=1}^{N_{MC}+1} I(d_j^2 \geq d_c^2)}{N_{MC} + 1}$$

sendo d_c^2 a distância obtida na amostra original e $I(d_j^2 \geq d_c^2)$ a função indicadora, que retorna 1 se $d_j^2 \geq d_c^2$ ou retorna 0, em caso contrário, para a j-ésima amostra Monte Carlo.

O julgamento, considerando um nível nominal de significância α, deverá confrontar o valor-p e α. As rotinas para a aplicação do teste de normalidade proposto foram feitas usando o programa estatístico R (R Development Core Team, 2012). Uma função para aplicar e outra para validação foram implementadas. O teste Monte Carlo de normalidade univariado baseado em distâncias foi denotado por TMCBD e o teste de normalidade Shapiro-Wilk por TNUSW.

2.2 Validação do teste

Duas etapas foram usadas para validar o teste. Na primeira, foram realizadas simulações sob H_0 dada em (10), ou seja, foram simuladas amostras de tamanho n da distribuição normal. Os dados foram simulados de uma distribuição normal padrão, sem perda de generalidade. O teste de Shapiro-Wilk, utilizado para fins de comparação, foi aplicado, a cada amostra simulada, utilizando os recursos do programa R, por meio da função *shapiro.test*. Em ambos os testes confrontou-se o nível de significância α pré-fixado com os valores-p em cada amostra.

Um número N de simulações Monte Carlo de validação foram realizadas sob H_0, sendo ambos os testes aplicados. A proporção de rejeições da hipótese nula de normalidade foi computada para o total de N simulações. Estes valores são estimativas do tamanho real do teste, que é o máximo (ou supremo) das probabilidades de se cometer o erro tipo I. Para avaliar o efeito da aleatoriedade resultante do processo de Monte Carlo foi aplicado um teste binomial exato para a hipótese nula de que o nível de significância do teste é igual ao valor nominal, ou seja, para a hipótese que o teste é exato (Oliveira e Ferreira, 2010).

Os níveis nominais de significância considerados foram iguais a 0,10, 0,05 e 0,01 e os tamanhos de amostras n iguais a 5, 10, 30, 100, 500, 5.000 e 10.000. O número de simulações Monte Carlo para a validação do teste N foi igual 2.000. O número de simulações Monte Carlo N_{MC} para a aplicação do teste foi também igual a 2.000.

Foram aplicados testes binomiais exatos, considerando o nível nominal de significância de 1%, para as hipóteses $H_0 : \alpha = 10\%$ versus $H_1 : \alpha \neq 10\%$, $H_0 : \alpha = 5\%$ versus $H_1 : \alpha \neq 5\%$ e $H_0 : \alpha = 1\%$ versus $H_1 : \alpha \neq 1\%$. Se a hipótese nula for rejeitada e os valores observados das taxas de erro tipo I forem considerados significativamente ($p < 0{,}01$) inferiores ao nível nominal, o teste deve ser considerado conservativo; se as taxas de erro tipo I forem consideradas significativamente ($p < 0{,}01$) superiores ao nível nominal, o teste deve ser considerado liberal; e se os valores observados das taxas de erro tipo I não diferirem significativamente ($p > 0{,}01$) do nível nominal, o teste deve ser considerado exato. Considerando que y representa o número de hipóteses nula de normalidade rejeitadas (sucesso) nas N simulações Monte Carlo para o nível de significância nominal α, então a estatística do teste é obtida, considerando a relação entre as distribuições F e binomial (Leemis

e Trivedi, 1996), com probabilidade de sucesso $p = \alpha$, por

$$F_c = \left(\frac{y+1}{N-y}\right)\left(\frac{1-\alpha}{\alpha}\right),$$

que, sob a hipótese nula, segue a distribuição F de Fisher-Snedecor com $\nu_1 = 2(N-y)$ e $\nu_2 = 2(y+1)$ graus de liberdade. Se $F_c \leq F_{0,005}$ ou $F_c \geq F_{0,995}$, a hipótese nula deve ser rejeitada no nível nominal de significância de 1%, em que $F_{0,005}$ e $F_{0,995}$ são quantis da distribuição F com ν_1 e ν_2 graus de liberdade.

O poder do teste proposto foi avaliado na segunda etapa de validação, sendo comparado com o do teste de normalidade de Shapiro-Wilk. Isso foi feito simulando N amostras considerando os mesmos tamanhos amostrais anteriormente definidos para a avaliação do erro tipo I. Também foram considerados os mesmos níveis nominais de significância já mencionados. Para avaliar o poder foram simuladas amostras sob H_1, ou seja, sob outra distribuição diferente da normal. Escolheu-se algumas distribuições de probabilidades comumente encontradas nas pesquisas.

Foi considerada inicialmente a distribuição t de *Student* com $\nu = 1$ e 30 graus de liberdade. Tal escolha foi feita devido a semelhança da sua forma com a da normal, ou seja, é uma distribuição simétrica centrada em 0. Se os graus de liberdade são pequenos, a distribuição se afasta da normal, embora com a mesma forma; se os graus de liberdade forem grandes, a distribuição se aproxima consideravelmente da normal, e espera-se que o poder do teste seja pequeno. Também foi escolhida a distribuição gama padrão, ou seja, com um único parâmetro, por ser uma distribuição assimétrica à direita. Escolheu-se uma gama com parâmetros $a = 0{,}5$ e $a = 1{,}5$. Para representar as distribuições assimétricas, também se escolheu a distribuição lognormal padrão. Finalmente, considerou-se uma distribuição beta com parâmetros $a = 1$ e $b = 1$, que corresponde a distribuição uniforme (0, 1). Essa escolha se deveu ao fato de a distribuição uniforme ser platicúrtica e simétrica. Assim, espera-se contemplar uma série de possibilidades na avaliação do poder em relação à forma das distribuições consideradas sob H_1.

Para o cômputo do poder, do mesmo modo que realizado para o erro tipo I, as taxas de rejeições de H_0 foram registradas nas N simulações Monte Carlo de cada configuração considerada. O teste Shapiro-Wilk também foi aplicado, como na primeira etapa de validação, para fins de comparação. A exceção se deu para amostras de tamanho $n = 10.000$, uma vez que o teste de Shapiro-Wilk é aplicável à amostras de tamanhos máximos de 5.000. Para suplantar essa dificuldade que ainda persiste foi construído o presente teste Monte Carlo, cuja restrição para lidar com grandes amostras é unicamente atribuída a memória disponível nos computadores. Todas as simulações foram realizadas no R, utilizando as funções de geração de amostras aleatórias das distribuições anteriormente mencionadas.

3 Resultados e discussão

Os resultados obtidos com a validação do teste Monte Carlo de normalidade univariado baseado em distâncias (TMCBD) e sua comparação com o teste de

normalidade Shapiro-Wilk (TNUSW) são avaliados e discutidos nesta seção. Para isso, a mesma foi dividida em duas partes, sendo a primeira subseção para avaliar o erro tipo I e a segunda, para avaliar o poder dos testes comparados neste trabalho. Considerando o erro tipo I, a comparação dos dois testes mencionados anteriormente foi feita para os níveis de significância α, fixados em 0,10 , 0,05 e 0,01, considerando os diferentes tamanhos de amostra.

Na subseção referente ao poder dos testes, as comparações foram feitas para os mesmos valores de significância nominais e mesmos tamanhos de amostra considerados na primeira subseção, apesar de as simulações terem sido feitas sob distribuições não-normais univariadas.

3.1 Erro tipo I

Um erro tipo I é cometido em um teste de hipóteses, ao se rejeitar a hipótese nula quando esta é verdadeira. A probabilidade de se cometer esse erro, se o teste for exato, é denotada por α e chamado de nível de significância ou tamanho do teste. A proporção de simulações do total de 2.000, em que os valores-p foram menores ou iguais ao nível de significância nominal adotado, serviu para estimar a probabilidade o erro tipo I real. Sendo assim, considera-se que o teste de normalidade proposto, controla adequadamente o erro tipo I, quando a taxa de erro empírica estimada estiver próxima do valor do nível de significância nominal. Tal proximidade é medida por um teste binomial, apresentado na seção de validação, minimizando assim o efeito dos erros de Monte Carlo e o caráter subjetivo desse julgamento.

Os resultados obtidos com as simulações, para os dois testes de normalidade comparados neste trabalho, estão apresentados na Tabela 1. Os dois testes apresentam, de modo geral, controle adequado de erro tipo I, com taxas equivalentes. Ademais, a aplicação do teste binomial aos resultados de ambos os testes (Tabela 1) permitiu que concluísse tanto o teste proposto como o de Shapiro-Wilk são considerados exatos, nos três níveis de significância nominais adotados e nas configurações adotadas. Deve-se chamar a atenção que a distribuição nula foi obtida por simulação Monte Carlo, o que permitiu que se definisse um teste exato por construção, como evidenciado pelos resultados obtidos. Entretanto, deve-se salientar que o uso de apenas uma amostra finita dos resultados possíveis da distribuição nula poderia ter acarretado uma imprecisão nos resultados esperados. Mas, haja vista os resultados observados, o número de simulações Monte Carlo de 2.000 foi suficiente para que isso não tenha ocorrido nos três níveis de significância adotados. Para níveis de significância nominais inferiores a 1%, sugere-se utilizar um número maior de simulações Monte Carlo, tanto na obtenção do valor-p do teste quanto na validação.

Chama a atenção o fato de que o teste proposto pode ser aplicado em uma amostra de tamanho $n = 10.000$, por exemplo, que é uma situação em que o teste de normalidade de Shapiro-Wilk não pode ser aplicado, em virtude de suas limitações computacionais. Esse fato revela uma das principais vantagens da presente proposta.

Table 1 - Taxas de erro tipo I dos testes de normalidade univariada TMCBD e TNUSW para os níveis nominais de significância de 10%, 5% e 1% em função do tamanho da amostra n

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCBD	TNUSW	TMCBD	TNUSW
5	0,1155[‡]	0,0985	0,0535	0,0465	0,0090	0,0075
10	0,0990	0,0855	0,0485	0,0450	0,0110	0,0065
30	0,1035	0,1070	0,0480	0,0500	0,0095	0,0100
100	0,0920	0,0955	0,0455	0,0495	0,0095	0,0080
500	0,1045	0,0975	0,0525	0,0525	0,0100	0,0120
5.000	0,0990	0,0890	0,0530	0,0445	0,0090	0,0120
10.000	0,1070	-	0,0510	-	0,0095	-

[‡] os resultados de ambos os teste em todos os casos não diferiram significativamente ($P > 0{,}01$) do valor nominal de significância correspondente.

3.2 Poder

O erro tipo II é cometido em um teste quando a hipótese nula não é rejeitada, sendo ela falsa. Denota-se por β a probabilidade de se cometer esse erro. O poder de um teste é o complemento, em relação à unidade, desta probabilidade, ou seja, é a probabilidade de se rejeitar a hipótese nula, quando ela é falsa. Deste modo, para se comparar o poder dos dois testes, TMCBD e TNUSW, calculou-se o percentual de rejeições da hipótese nula, quando amostras de distribuições não-normais foram simuladas. Entre os dois testes, aquele que detectar uma maior quantidade de amostras não-normais é o que tem maior poder. Os valores obtidos nas simulações foram apresentados em tabelas para cada distribuição considerada, a fim de facilitar a comparação.

Considerando a distribuição não-normal *t-Student* com grau de liberdade $\nu = 1$, os valores de poder para o TMCBD e o TNUSW são apresentados na Tabela 2. Observa-se, para este tipo de distribuição não normal, que o TNUSW apresenta maior poder do que o teste proposto na maioria dos casos em pequenas amostras ($n \leq 10$), embora as diferenças não sejam muito expressivas. À medida que o tamanho das amostras cresce, o desempenho de ambos os testes tende a se igualar, sendo que ambos os testes atingiram 100% de poder em amostras a partir se $n = 100$, ou seja, detectam o total das amostras não-normais simuladas para todos os níveis de significância nominais adotados. Nas amostras de tamanho $n = 10.000$, apenas o TMCBD foi aplicado. Em amostras de tamanhos iguais ou superiores a 30, o TMCBD igualou-se ao TNUSW e em amostras inferiores a 30, o TNUSW foi superior. Nesse último caso, quanto menor for α, melhor é o desempenho relativo do TNUSW em relação ao TMCBD.

Os desempenhos em relação ao poder dos TMCBD e TNUSW considerando a distribuição não-normal *t-Student* com grau de liberdade $\nu = 30$ para $\alpha = 10\%$, $\alpha = 5\%$ e $\alpha = 1\%$ são mostrados na Tabela 3. Para este tipo de distribuição, o

Table 2 - Poder dos testes para os níveis nominais de significância de 10%, 5% e 1% relativo à distribuição t com $\nu = 1$ grau de liberdade, considerando diferentes tamanhos de amostras

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCDB	TNUSW	TMCBD	TNUSW
5	0,3485	0,3675	0,2670	0,281	0,1235	0,1655
10	0,6565	0,6605	0,5825	0,6005	0,4395	0,4860
30	0,9675	0,9750	0,9555	0,9640	0,9230	0,9265
100	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
500	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5.000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10.000	1,0000	-	1,0000	-	1,0000	-

teste proposto apresenta maior poder em relação ao TNUSW praticamente para todos tamanhos de amostras. À medida que se aumenta o tamanho da amostra, o desempenho do TMCBD se amplia sua vantagem em relação ao TNUSW. Para amostras de tamanho $n = 5.000$, essa superioridade oscila em torno de 5 pontos percentuais em poder para todos os níveis nominais de significância.

Em alguns casos, nas pequenas amostras, os valores de poder estão muito próximos dos valores nominais de significância. Isso ocorre devido à semelhança da distribuição t de *Student* com $\nu = 30$ com a distribuição normal. Convém destacar, ainda, o fato de que as distribuições simuladas nesses dois primeiros casos são simétricas, que teoricamente é uma condição que a hipótese nula de normalidade, falsa nesse caso, é mais difícil de ser rejeitada por um teste de normalidade. O TMCBD apresentou excelente desempenho para amostras de tamanhos 10.000.

Table 3 - Poder dos testes para os níveis nominais de significância de 10%, 5% e 1% relativo à distribuição t com $\nu = 30$ graus de liberdade

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCBD	TNUSW	TMCBD	TNUSW
5	0,1000	0,0920	0,0510	0,0375	0,0120	0,0080
10	0,1115	0,0990	0,0590	0,0545	0,0140	0,0120
30	0,1270	0,1205	0,0650	0,0705	0,0160	0,0170
100	0,1460	0,1375	0,0920	0,0810	0,0285	0,0270
500	0,2530	0,2345	0,1595	0,1520	0,0590	0,0575
5.000	0,7790	0,7210	0,6935	0,6350	0,4645	0,4370
10.000	0,9560	-	0,9235	-	0,8085	-

Os valores de poder dos TMCBD e TNUSW para a distribuição gama com parâmetro 0,5, nos três níveis de significância nominais adotados, estão apresentados na Tabela 4. O desempenho de poder do TMCBD é expressivamente superior em pequenas amostras ($n \leq 10$) para todos os níveis nominais de significância α. Em

amostras de tamanho $n = 5$, para os três níveis significâncias nominais adotados, a diferença entre os poderes dos dois testes a favor do TMCBD foi de maior magnitude do que a que ocorreu para $n = 10$. Em amostras a partir de 30, ambos os testes ou apresentaram poder igual a 100% ou muito próximo de 100% e similar, isto é, ambos detectaram praticamente a totalidade das amostras não-normais simuladas.

Table 4 - Poder dos testes para os níveis nominais de significância de 10%, 5% e 1% relativo à distribuição gama com $a = 0{,}5$

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCBD	TNUSW	TMCBD	TNUSW
5	0,5575	0,4290	0,4195	0,2985	0,2075	0,1375
10	0,8975	0,8220	0,8205	0,7390	0,6020	0,5105
30	1,0000	1,0000	1,0000	0,9995	0,9975	0,9970
100	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
500	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5.000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10.000	1,0000	-	1,0000	-	1,0000	-

Na tabela 5, os valores de poder são apresentados considerando a distribuição não normal univariada gama de parâmetro 1,5, relativos aos três níveis de significância nominais adotados. O desempenho de poder do TMCBD mostra-se expressivamente maior do que o desempenho do TNUSW em amostras de tamanho $n = 5$, $n = 10$ e $n = 30$. Para amostras de tamanhos $n = 100$, $n = 500$ e $n = 5.000$ o poder dos dois testes é de 100%. Somente o poder do TMCBD foi estimado para $n = 10.000$, que em todos os níveis de significância foi igual a 100%.

Table 5 - Poder dos testes para os níveis nominais de significância de 10%, 5% e 1% relativo à distribuição gama com $a = 1{,}5$

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCBD	TNUSW	TMCBD	TNUSW
5	0,2755	0,2010	0,1705	0,1215	0,0560	0,0445
10	0,5530	0,4215	0,4185	0,3110	0,1885	0,1400
30	0,9625	0,9215	0,9155	0,8590	0,7425	0,6640
100	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
500	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10000	1,0000	-	1,0000	-	1,0000	-

Para a distribuição lognormal padrão, os desempenhos de poder dos dois testes estão apresentados na Tabela 6, para os três níveis de significância nominais adotados. Nota-se novamente uma expressiva superioridade de desempenho de poder do TMCBD em relação ao TNUSW em amostras de tamanhos $n = 5$ e $n = 10$. Para amostras de tamanho $n = 30$, o desempenho de poder do TMCBD

continua superior, porém com valores muito próximos. Em amostras de tamanhos $n = 100$, $n = 500$ e $n = 5.000$ ambos os testes apresentam 100% de poder.

A lognormal é assimétrica à direita. Em pequenas amostras, nesse tipo de distribuição, verificou-se que, pelo menos nas distribuições consideradas neste trabalho, o desempenho do TMCBD foi superior, em virtude das grandes diferenças de poder encontradas. Os dois testes apresentam desempenhos equivalentes em grandes amostras, o que indica equivalência assintótica.

Table 6 - Poder dos testes para os níveis nominais de significância de 10%, 5% e 1% relativo à distribuição lognormal padrão

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCBD	TNUSW	TMCBD	TNUSW
5	0,4485	0,3670	0,3255	0,2520	0,1575	0,1090
10	0,7940	0,7125	0,6985	0,6220	0,4825	0,4250
30	0,9995	0,9950	0,9960	0,9915	0,9785	0,9700
100	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
500	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5.000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10.000	1,0000	-	1,0000	-	1,0000	-

Na Tabela 7 estão apresentados os desempenhos de poder para a distribuição beta (1,1). Observa-se que nesse caso houve uma inversão do desempenho de poder dos dois testes em relação ao que vinha ocorrendo anteriormente, para amostras de tamanhos $n = 5$, $n = 10$, $n = 30$ e 100. Em amostras de tamanhos $n = 500$ e $n = 5.000$, ambos os teste apresentam desempenhos equivalentes, detectando a totalidade das amostras não-normais simuladas. A distribuição beta (1,1) é na verdade a U(0,1), que é platicúrtica e simétrica. Mesmo assim, o teste proposto apresentou desempenho satisfatório quando comparado com o TNUSW em amostras de menores tamanhos. Em amostras grandes, os dois testes se equivaleram e com desempenho ótimo, pois atingiram 100% de poder.

Table 7 - Poder dos testes para os níveis nominais de significância de 10%, 5% e 1% relativo à distribuição beta - (1,1)

	$\alpha = 0{,}10$		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n	TMCBD	TNUSW	TMCBD	TNUSW	TMCBD	TNUSW
5	0,1215	0,1255	0,0585	0,0605	0,0135	0,0125
10	0,1235	0,1590	0,0505	0,0805	0,0095	0,0150
30	0,3585	0,5680	0,1780	0,3760	0,0215	0,0835
100	0,9925	1,0000	0,9625	0,9980	0,7225	0,9510
500	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10000	1,0000	-	1,0000	-	1,0000	-

Novas distribuições sob H_1 devem ser consideradas em trabalhos futuros, de modo que se faça uma investigação mais ampla sobre os desempenhos deste teste proposto, comparado com o teste de normalidade Shapiro-Wilk.

Conclusões

Os dois testes apresentaram resultados equivalentes em relação ao controle de erro tipo I, uma vez que tanto para o TMCBD quanto para o TNUSW, os valores dos tamanhos empíricos dos dois testes não diferiram, exceto pelo erro de Monte Carlo, dos valores nominais de significância, na totalidade das configurações simuladas, indicando que ambos os testes são exatos.

Em relação ao desempenho de poder, o TMCBD mostra-se mais poderoso do que o TNUSW na maioria das distribuições verificadas, nos níveis nominais de significância fixados. O teste proposto é absolutamente intuitivo pela clareza de sua fundamentação teórica. Além disso, o tempo de obtenção dos resultados no teste proposto é pequeno, quando usado um processador de bom desempenho. O TMCBD pode ser usado em amostras superiores a 5.000, que uma de suas principais virtudes.

PEREIRA FILHO, N. A. ; FERREIRA, D. F. Monte Carlo test of normality based on distance. *Rev. Bras. Biom.,* São Paulo, v.30, n.3, p.401-416, 2012.

▪ *ABSTRACT: The normal probability distributions describing the behavior of many real-life phenomena in various fields of science. When one considers a random sample of a population, in the univariate case, it is common to assume that the data or residuals of the model are normally distributed. Graphs such as histograms and Q-Qplots are quite effective, but subjective, to check the normality. However, this is not enough to assure the normality of the data or the residuals of some fitted model. There are several tests of normality in the literature. Among them, the Shapiro-Wilk is considered to pursue optimal properties. However, this test is computationally applicable to a number of observations up to* 5,000. *This paper aims to propose a univariate normality test, based on distances between the observed values of the sample order statistics and the expected values of the standard normal order statistics, that can be applied to any sample sizes with no theoretical restrictions. The distribution of the test statistic was obtained by Monte Carlo simulation. The results of power and type I error rates, allow the conclusion that the proposal test is generally more efficient than the Shapiro-Wilk test and does not have the practical limitation of being restricted to size up to* 5,000.

▪ *KEYWORDS: Normality; distances; power; Shapiro-Wilk.*

References

FILLIBEN, J.J. The Probability plot correlation coefficient test for normality. *Tecnometrics*, Washigngton, v.17, n.1, 1975, p.11-117.

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

OLIVEIRA, I.R.C.; FERREIRA, D.F. Multivariate extension of chi-squared univariate normality test. *J. Stat. Comput. Simul.*), London, v.80, n.5, p.513-526, 2010.

R DEVELOPMENT CORE TEAM. *R*: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. Disponível em:¡http://www.R-project.org¿.

ROMÃO, X.; DELGADO, R.; COSTA, A. An empirical power comparison of univariate goodness-of-fit of normality.*J. Stat. Comput. Simul.*), London, v.80, n.5, 2010, p.545-591.

ROYSTON, J.P. An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Stat. - J. R. Stat. Soc. - Ser. C*, London, v. 31, n. 2, p.115-124, 1982.

ROYSTON, J.P. Approximating the Shapiro-Wilk's W-test for non-normality. *Stat. Comput.*, London, v. 2, n. 1, p.117-119, 1992.

ROYSTON, J.P. A toolkit for testing for non-normality in complete and censored samples. *Statistician*, London, v. 42, n. 1, p.37-43, 1993.

ROYSTON, J.P. A remark on algorithm AS 181: the W-test for normality. *Appl. Stat. - J. R. Stat. Soc. - Ser. C*, London, v. 44, n. 4, p.547-551, 1995.

SHAPIRO, S. S.; WILK, M.B. An Analysis of Variance Test for Normality (complete samples), *Biometrika*, London, v.52,n.3-4, p.591-611, 1965.

THODE, H.JR. *Testing for normality*. New York: Marcel Decker, 2002. 479p.

YAZICI, B.; YOLACAN, S. A comparison of various tests of normality. *J. Stat. Comput. Simul.*, London, v.80, n.2, p.175-183, 2007.

Recebido em 23.10.2012.

Aprovado após revisão em 19.02.2013.