

MODIFICAÇÃO DO TESTE DE NORMALIDADE DE SHAPIRO-WILK MULTIVARIADO DO SOFTWARE ESTATÍSTICO R

Roberta Bessa Veloso ¹, Daniel Furtado Ferreira ², Eric Batista Ferreira ³

INTRODUÇÃO

A inferência estatística consiste em fazer generalizações sobre uma população com base nos dados amostrais. O problema de se inferir, a partir de dados mensurados pelo pesquisador, sobre os processos ou fenômenos físicos, biológicos ou sociais, que não se pode diretamente observar, é uma realidade constante. Neste contexto, todo o processo de inferência resume-se na estimação por intervalo e testes de hipóteses de parâmetros populacionais. Presume-se que os dados sejam provenientes de uma população normal devido ao fato de seu modelo populacional apresentar-se coerente a muitos fenômenos aleatórios.

Para a realização de inferências no caso multivariado, seguem-se os mesmos princípios do caso univariado. Portanto, é de grande importância a normalidade dos dados, que generalizada pra muitas dimensões é conhecida por distribuição normal multivariada.

A identificação de normalidade nos dados, seja no caso univariado ou multivariado, em geral é feita por meio de gráficos, entretanto a simples constatação via gráficos não é suficiente, principalmente no caso multivariado e, especificamente nas situações de muitas variáveis. Sendo assim, a utilização de testes estatísticos como meio de inferir sobre a normalidade é necessária.

O procedimento de Royston (1983) prevê a estimação da estatística W de Shapiro-

¹Doutoranda em Estatística e Exp. Agrop., DEX/UFLA, bolsista FAPEMIG, e-mail: bessaveloso@yahoo.com.br.

²Professor Associado I, DEX/UFLA, bolsista CNPq, e-mail: danielff@ufla.br.

³Pós-doutorado, Departamento de Ciências Exatas (DEX/UFLA), bolsista CNPq, e-mail: ericbferreira@netscape.net.

Wilk para cada uma das variáveis, sendo a estatística final do teste baseada na soma dos seus valores. É utilizada uma transformação da estatística e a correlação entre as variáveis é utilizada para obter os graus de liberdade da distribuição de Qui-quadrado resultante (ROYSTON, 1983).

Uma forma de avaliar o desempenho de um teste é mensurar tanto as taxas de erro tipo I, em diferentes condições da hipótese nula de normalidade, quanto o poder do teste, simulando amostras sob a hipótese alternativa de não-normalidade (neste caso, o poder). Um teste ideal, apesar de não existir, seria aquele que não rejeitasse para nenhuma amostra observada a hipótese nula verdadeira e rejeitasse 100% das vezes as hipóteses nulas falsas. Como isso não ocorre em situações reais, busca-se um teste que mantenha as taxas de erro tipo I menores ou iguais a um valor nominal de probabilidade escolhido e que tenha o maior poder possível.

O software estatístico R (R DEVELOPMENT CORE TEAM, 2006) tem tido grande impacto no meio científico, e por ter código fonte aberto, tem recebido contribuições de pesquisadores de todo o mundo. A função *mshapiro.test* do pacote *mvnrmtest* possibilita ao usuário aplicar o teste de normalidade multivariada de Shapiro-Wilk, que não trata da extensão multivariada de Royston (1983). Esta função é baseada na generalização multivariada do teste proposto por Domanski (1998), que se baseia em buscar uma combinação linear das p variáveis originais e aplicar o teste de Shapiro-Wilk nesta nova variável.

A motivação para este trabalho surgiu a partir do interesse de avaliar o desempenho do teste multivariado de normalidade de Shapiro-Wilk implementado no R. Assim, foi proposto este trabalho objetivando-se comparar o desempenho do teste multivariado de normalidade de Shapiro-Wilk com a modificação proposta utilizando simulação Monte Carlo.

METODOLOGIA

Duas estratégias foram consideradas neste trabalho. A primeira teve o intuito de avaliar as taxas de erro tipo I dos teste de normalidade Shapiro-Wilk multivariado do

software R (MSWR) e com a modificação proposta (MSWM). A segunda foi delineada para avaliar o poder dos testes. Em ambos os casos, foi usada simulação Monte Carlo. Em cada simulação foram aplicados os testes de normalidade em um nível nominal pré-estabelecido de significância, sendo verificado se a hipótese nula foi ou não rejeitada. Este processo, em cada caso, foi repetido 10.000 vezes e a proporção de decisões incorretas no primeiro caso é a taxa de erro tipo I empírica e, no segundo caso, a proporção de decisões corretas é o poder empírico. Os valores da taxa de erro tipo I empírica foram comparados com o valor nominal por meio de um intervalo de confiança para proporções. Também foram comparadas as taxas de erro e o poder dos dois testes aplicados.

RESULTADOS E DISCUSSÃO

Modificação proposta

Primeiramente, a matriz de dados observados \mathbf{Y} , foi escrita da seguinte forma

$$\mathbf{Y} = \frac{\sum_{i=1}^p \mathbf{X} \mathbf{b}_i}{\sqrt{\lambda_i}} = \sum_{i=1}^p \begin{pmatrix} \frac{\mathbf{X}_1^t \mathbf{b}_i}{\sqrt{\lambda_1}} \\ \frac{\mathbf{X}_2^t \mathbf{b}_i}{\sqrt{\lambda_2}} \\ \vdots \\ \frac{\mathbf{X}_n^t \mathbf{b}_i}{\sqrt{\lambda_p}} \end{pmatrix},$$

em que o vetor \mathbf{b}_i e λ_i foram definidos por

$$\lambda_i = \max_{\mathbf{b}_i} \frac{\mathbf{b}_i^t \mathbf{S} \mathbf{b}_i}{\mathbf{b}_i^t \mathbf{b}_i},$$

sendo $\mathbf{S} = \mathbf{W}/(n-1)$ e o vetor \mathbf{b}_i correspondente são obtidos da solução do polinômio característico $(\mathbf{S} - \lambda_i \mathbf{I}) \mathbf{b}_i = \mathbf{0}$, para $i = 1, 2, \dots, p$. Logo, λ_i e \mathbf{b}_i são o autovalor e o autovetor de \mathbf{S} , com $\mathbf{b}_i^t \mathbf{b}_i = 1$.

Em seguida, é definida a combinação linear $\mathbf{Z}_i = \mathbf{X} \mathbf{b}_i$ e o vetor $\mathbf{Y}_i = \sum_i^p \mathbf{Z}_i$ é submetido ao teste de normalidade de Shapiro-Wilk univariado, reportando o valor da

estatística W e o valor- p .

Erro tipo I e poder

Na Tabela 1 estão apresentadas as taxas de erro tipo I para os testes, (MSWM) e (MSWR) sob normalidade multivariada em função de ρ , p e n_i . O intervalo de 99% de confiança para o valor nominal de 5% é [0,0457; 0,0542] e os resultados obtidos foram confrontados com este intervalo.

O MSWM controlou o erro tipo I em todas as situações simuladas considerando o nível nominal de 5% de significância. O MSWR não controlou o erro tipo I em nenhuma das situações consideradas nesse trabalho. Este teste foi extremamente liberal apresentando taxas de erro superiores ao nível nominal considerado.

TABELA 1: Taxas de erro tipo I para os dois testes, considerando normalidade multivariada, valor nominal de 5%, em função dos diferentes tamanhos amostrais (n_i), do número de variáveis p e da correlação ρ .

n	ρ	MSWM	MSWR
5	0	0,0478	0,3108*
5	0,1	0,0464	0,3173*
5	0,5	0,0478	0,3108*
5	0,9	0,0475	0,3143*
20	0,5	0,0467	0,1343*
30	0,5	0,0520	0,1217*
60	0,5	0,0543	0,1092*
100	0,5	0,0455	0,1077*
200	0,5	0,0480	0,0989*
300	0,5	0,0490	0,0942*
400	0,5	0,0497	0,0940*
500	0,5	0,0470	0,0914*
750	0,5	0,0528	0,0944*
1000	0,5	0,0550	0,0972*
2000	0,5	0,0503	0,0820*

*Taxas de erro tipo I não pertencentes ao intervalo [0,0457; 0,0542]; ou seja, acima do nível nominal de 5%.

A correlação teve pouca ou nenhuma influência na taxa de erro de tipo I. Por exemplo, para $n = 5$, na Tabela 1, percebe-se que a correlação não alterou as taxas de erro tipo I dos testes. Por isso para os demais tamanhos da amostra foram apresentados os resultados para $\rho = 0,5$.

À medida que o número de variáveis, p aumentou de 2 para 10, o teste MSWM manteve o erro tipo I sob controle ao contrário do MSWR que apresentou taxas de erro igual a 100% (resultados não apresentados).

Para a avaliação do poder destes testes sob, H_1 , optou-se por fixar a correlação em $\rho = 0,5$ por não afetar a performance dos testes. Os resultados simulados permitiram detectar uma considerável superioridade do teste MSWM em relação ao teste MSWR.

CONCLUSÕES

O teste de normalidade de Shapiro-Wilk multivariado modificado neste trabalho obteve êxito quanto ao controle do erro tipo I e apresentou alto poder. O teste implementado no R apresentou taxas de erro tipo I muito elevadas e não deve ser recomendado.

REFERÊNCIAS BIBLIOGRÁFICAS

DOMANSKI, C. Własności testu wielowymiarowej normalności shapiro-wilka i jego zastosowanie.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

ROYSTON, J. P. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W . **Applied Statistics - Journal of the Royal Statistical Society - Series C**, 32(2):121–133, 1983b.