DEFINIÇÃO DO TAMANHO AMOSTRAL USANDO SIMULAÇÃO MONTE CARLO PARA O TESTE DE NORMALIDADE BASEADO EM ASSIMETRIA E CURTOSE. II. ABORDAGEM MULTIVARIADA

ANDRÉA CRISTIANE DOS SANTOS[1]
DANIEL FURTADO FERREIRA[2]

RESUMO – Realizou-se este trabalho com o objetivo de determinar um tamanho amostral ótimo para as estatísticas univariadas de assimetria e curtose (z_1 e z_2) que, neste caso, foram adaptadas para o caso multivariado, e as estatísticas multivariadas de assimetria e curtose (K_1 e K_2) com base em simulação. Foram geradas diferentes funções densidade probabilidade multivariadas via método de Monte Carlo para avaliar a taxa de erro tipo I e o poder do teste para os valores nominais de 5% e 1%. Foram avaliadas as situações com p=2, 3, 4 e 5 variáveis, com diferentes estruturas de correlação. Para o caso multivariado, as diferentes estruturas de correlação não afetaram o poder e a taxa de erro tipo I dos testes; a estatística K_1 é adequada para uso a partir de n≥50 para valores nominais de significância de 5% ou 1%; a estatística K_2 é assintoticamente adequada para os testes de desvios de curtose para n≥100, independentemente dos valores nominais da significância. Pode-se concluir que as estatísticas de assimetria, em geral, são mais poderosas do que as de curtose, mas os testes da hipótese nula de normalidade multivariada devem considerar tanto os testes de desvios de assimetria como os de curtose conjuntamente, como sugerido no caso univariado.

TERMOS PARA INDEXAÇÃO: Assimetria, curtose, teste normalidade multivariado, testes assintóticos.

SAMPLE SIZE DEFINITION USING MONTE CARLO SIMULATION FOR THE NORMALITY TEST BASED ON SKEWNESS AND KURTOSIS. II. MULTIVARIATE APPROACH

ABSTRACT – This work aimed to determine optimum sample size for the univariate skewness and kurtosis statistics (z_1 and z_2) adapted to multivariate situation and for the multivariate skewness and kurtosis statistics (K_1 and K_2) statistics based on simulation. Different probability density functions, univariate and multivariate, were generated by Monte Carlo simulation method to evaluate the type I error rates and the power of the tests. The simulations were done adopting the nominal significance level of 5% and 1%. Situations with p=2, 3, 4 and 5 variables with different correlation structures were evaluated in the case of multivariate distributions. The results showed that K_1 statistics is adequate for n≥50, at nominal levels of significance of 5 or 1%; different correlation structured do not affect the power and the type I error rates, the K_2 statistics is asymptotically appropriate for kurtosis deviation tests for n≥100, independently of the nominal values of the significance. The skewness statistics, in general, were shown to be more powerful than those of kurtosis, however, the null hypothesis tests of normality must consider both tests jointly, as suggested in the univariate case.

INDEX TERMS: Skewness, kurtosis, test for normality, type I error rates and power of the test.

1. Mestre em Estatística e Experimentação Agropecuária, Professor de Estatística da FUOM, Formiga, MG, andrea@ufla.br.
2. Dr. em Genética e Melhoramento de Plantas, Professor Adjunto III do Departamento de Ciências Exatas da UNIVERSIDADE FEDERAL DE LAVRAS/UFLA, Caixa Postal 37 – 37200-000 – Lavras, MG, danielff@ufla.br, Bolsista CNPq.

INTRODUÇÃO

Verificar a suposição de normalidade multivariada é imprescindível para a utilização de alguns métodos estatísticos. No entanto, no caso multivariado, há uma grande dificuldade para verificar essa suposição. Pouca informação existe na literatura abordando testes estatísticos de normalidade multivariada. Segundo Bock (1975), uma alternativa para verificar a suposição de normalidade multivariada é pela verificação da normalidade da distribuição marginal univariada. Sabe-se que a existência de uma distribuição normal multivariada implica em marginais normalmente distribuídas. Contudo, não se pode garantir que a distribuição conjunta de duas variáveis normais univariadas sejam uma normal multivariada. Dessa forma, torna-se inviável a verificação da hipótese de normalidade via marginais.

Para verificar a suposição de normalidade multivariada, é possível usar testes baseados nos desvios de assimetria e curtose. Esses testes são baseados em estatísticas assintóticas. Na literatura, são escassas as informações dos tamanhos amostrais ideais para a sua utilização. Usando simulação, Machado (1983) avaliou duas estatísticas, bastante complexas, baseadas nos coeficientes de assimetria e curtose, para testar normalidade multivariada. Pelos estudos de taxas de erro tipo I, para diferentes valores nominais de significância, concluiu-se que o tamanho $n \geq 25$ garante uma boa aproximação assintótica dessas estatísticas. Mudholkar et al. (1992) desenvolveram um teste p-variado, denotado Z_p, baseado na distância de Mahalanobis para testar normalidade multivariada. As taxas de erro tipo I e poder do teste foram avaliadas via simulação Monte Carlo. O teste atinge poder razoável com amostra n>50.

Os estimadores univariados e multivariados dos coeficientes de assimetria e curtose foram apresentados em Mardia (1970).

No caso univariado, é aconselhável destacar que as estatísticas Z_1 e Z_2 foram adaptadas para o caso multivariado por:

$$Z_1^2 = \frac{\hat{\beta}_{1,p}(n+1(n+3)}{6(n-2)} \sim \mathrm{X}^2 com \frac{p(p+1)(p+2)}{6} graus\ de\ liberdade$$

e

$$Z_2 = \left(\hat{\beta}_{2,\mathrm{p}} - 3 + \frac{6}{n+1}\right)\sqrt{\frac{(n+1)^2(n+3)(n+5)}{24(n-2)(n-3)}} \sim N(0,1)$$

O objetivo quando da realização deste trabalho foi determinar um tamanho amostral ótimo para as estatísticas univariadas (Z_1 e Z_2), que neste caso, foram adaptadas para o caso multivariado, e as estatísticas multivariadas (K_1 e K_2) com base em simulação.

METODOLOGIA

Para simular amostras multivariadas, foram considerados os mesmos tamanhos amostrais do caso univariado, n = 5(5)100(50)500. As situações avaliadas continham p=2, 3 e 5 variáveis. Cada situação foi simulada 5.000 vezes. As distribuições consideradas foram: normal multivariada, uniforme multivariada, distribuição multivariada gerada a partir da distribuição exponencial e a partir da distribuição de qui-quadrado. Foram consideradas diferentes estruturas de covariância (ρ: 0; 0,5; 0,9; -0,5).

Para medir as taxas de erro tipo I, foram consideradas as situações sob f.d.p. normais multivariadas, para os níveis nominais de 5% e 1% e computadas as proporções de rejeições (erros tipo I). A validade do teste aplicado foi verificada comparando-se o valor nominal adotado e a taxa empírica computada. O poder do teste foi avaliado considerando situações sob f.d.p. não normais multivariadas, computando-se as proporções em que a hipótese de normalidade multivariada não foi rejeitada.

A estrutura da matriz de covariância populacional considerada foi:

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{P-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{P-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{P-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{P-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{P-1} & \rho^{P-2} & \rho^{P-3} & \rho^{P-4} & \cdots & 1 \end{bmatrix} \sigma^2$$

em que ρ é um valor pertencente ao intervalo $]-1, 1[$; σ^2 é uma constante maior que zero.

Foram considerados os seguintes valores de ρ: 0; 0,5; 0,9; -0,5. Sem perda de generalidade, foi consi-

derada a distribuição normal multivariada com $\sigma^2 = 1$ e vetor de médias $\underset{\sim}{\mu} = \underset{\sim}{0}$.

Primeiramente, gerou-se um vetor $z_{p\times 1}$ de observações normais padronizadas (ou de outra distribuição sob estudo), cujos elementos são independentes. Para gerar tal vetor, usou-se o teorema da probabilidade integral. Para gerar a distribuição com covariância Σ, essa foi fatorada para obtenção do fator Cholesky por: $\Sigma = FF'$.

Fazendo $\underset{\sim}{\mu} = \underset{\sim}{0}$, obteve-se o vetor $\underset{\sim}{x}$ da seguinte forma:

$$\underset{\sim}{x} = Fz + \underset{\sim}{\mu}$$

o qual possui média $\underset{\sim}{0}$ e covariância Σ.

O processo de transformação linear singular aplicada a um vetor normal multivariado garante que não há alteração na distribuição (Johnson & Wichern, 1998). O processo foi repetido até que fossem geradas tantas observações multivariadas ($\underset{\sim}{x}$) quanto fosse o tamanho da amostra.

RESULTADOS E DISCUSSÃO

Para avaliar a taxa de erro tipo I, foi utilizado o caso sob distribuição normal multivariada, com p=2, 3, 4 e 5. As correlações utilizadas foram 0; 0,5; 0,9; -0,5. Os valores nominais adotados foram 5% e 1%. Na Figura 1 foram apresentadas as 4 estatísticas, considerando as diferentes estruturas de correlação. Como de uma forma geral observou-se que as diferentes estruturas de correlação não afetaram as taxas de erro tipo I e o poder dos testes, elas foram omitidas nas Figuras de 2 a 8. Esse aspecto é importante para o usuário desses testes pelo fato de não ser necessário nenhum tipo de preocupação com a natureza da estrutura de correlação quando da aplicação prática dos mesmos.

Por meio da Figura 1(a), da distribuição normal bivariada com α=5%, verificou-se que z_1 não controlou adequadamente a taxa de erro tipo I, superando o valor nominal estabelecido para n<150. A estatística z_1 aproximou-se do valor nominal com amostras n≥150. A estatística K_1, por outro lado, mostrou-se conservadora para n<75, começando a oscilar em torno do valor nominal a partir desse tamanho amostral, ou seja, n≥75. Para o valor nominal de significância de 1% (Figura 1b), a estatística K_1 apresentou poder inferior ao valor nominal para n<25 e z_1 apresentou valores superiores ao valor nominal para n<250. A partir desses valores de n, houve uma convergência dessas estatísticas para o valor nominal de 1%. As estatísticas z_2 e K_2, com α=5% mostradas na Figura 2(a), mantiveram-se aquém do valor nominal estabelecido. Isso ocorreu com n<50 para z_2 e com n<500 para K_2, e a estatística K_2 foi a mais rigorosa. Acima desses valores de n, as aproximações assintóticas podem ser consideradas adequadas. Por meio da Figura 2(b), verificou-se que a estatística K_2 foi mais rigorosa com n<50 e aproxima-se do valor nominal de 1% com n superior a esse valor. Para esses tamanhos amostrais e valor nominal, pode-se considerar que esta estatística controla a taxa de erro tipo I. A estatística z_2, por outro lado, só começou a controlar a taxa de erro tipo I para valores de n superiores a 500, e no caso contrário (n≤500), manteve-se acima do valor nominal fixado.

Para avaliar o poder do teste, foram consideradas distribuições cujas f.d.p's são não normais. Os resultados obtidos nas situações em que a distribuição multivariada foi gerada, a partir de uma distribuição de qui-quadrado e a partir de uma distribuição exponencial, foram similares. Foi apresentado apenas o caso p=2, para a distribuição de qui-quadrado, uma vez que utilizando p=3, 4 e 5, obtiveram-se resultados semelhantes para as duas distribuições. Destacou-se apenas o fato de que com o aumento das variáveis, a estatística z_1 aumentou o poder consideravelmente para amostras n<20. Esse resultado foi facilmente explicado considerando o fato de z_1 não ter controlado a taxa de erro tipo I adequadamente, em geral superestimando o valor de α. Verificou-se (Figuras 3a e 3b) que z_1 superou em poder a estatística K_1, o que, conforme comentado anteriormente, já era esperado. Quando n≥20, as estatísticas atingiram um elevado poder para α=5% e n≥30 para α=1%, tendendo a se igualar com n≥50.

Considerando as estatísticas relativas à curtose, verificou-se que a estatística Z_2 superou em poder a estatística K_2 (Figuras 4ª e 4b). Essas estatísticas atingiram um valor elevado de poder quando n≥70 para α=5% e n≥80 para α=1%. Essas estatísticas tendem a se igualar em poder quando n>100 para α=5% e n>130 para α=1%. Pelas mesmas razões mencionadas anteriormente, justificou-se o maior poder da estatística Z_2 sobre a K_2.

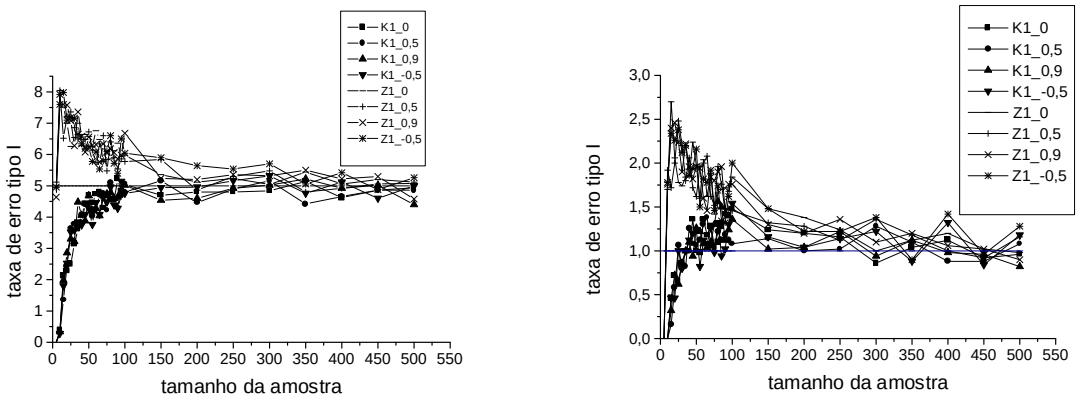


FIGURA 1 – Taxa de erro tipo I considerando a distribuição normal bivariada, 5% (a), 1% (b).

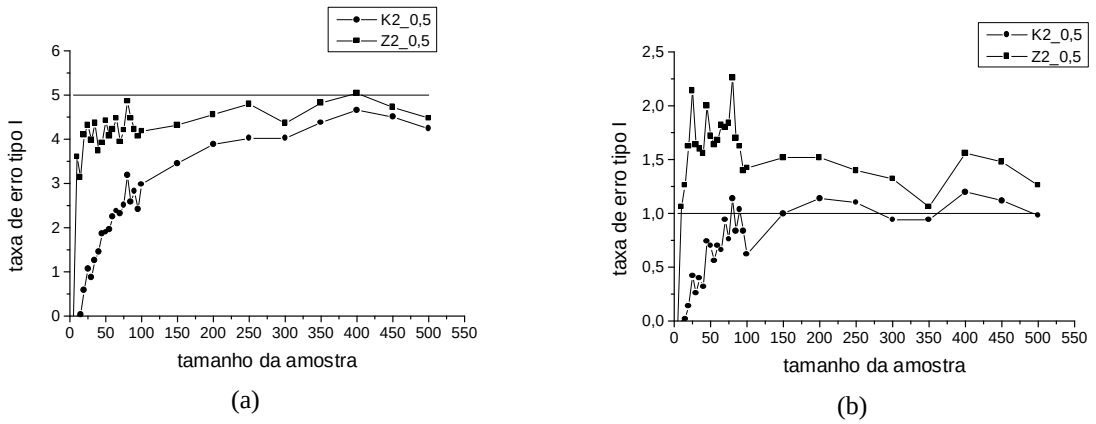
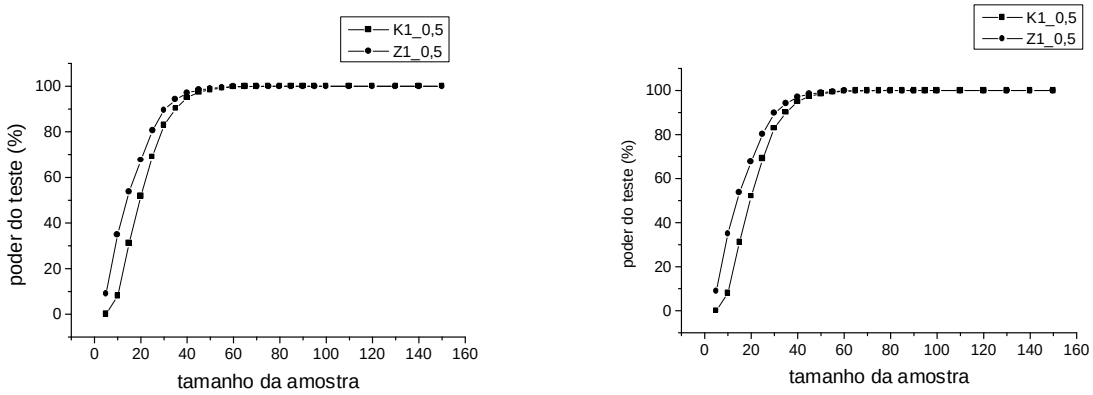


FIGURA 2 – Taxa de erro tipo I considerando a distribuição normal bivariada, 5% (a) e 1% (b).

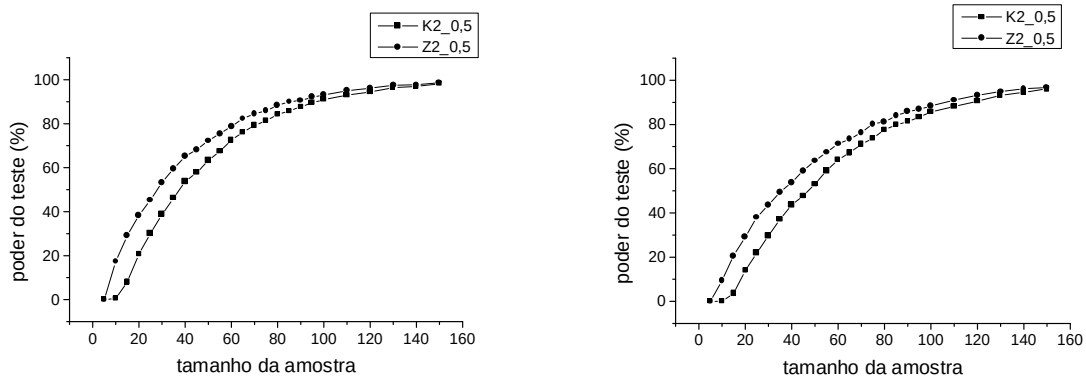
Para a distribuição uniforme bivariada, como mostra a Figura 5, verificou-se que a estatística K_1 não apresentou poder algum em detectar o desvio de normalidade; já a estatística Z_1 apresentou um baixo poder em detectar esse desvio para n≤20. Esse fato só pode significar uma inadequação dessa aproximação assintótica para valores pequenos de n, como esses. As estatísticas K_2 e Z_2 (Figuras 6a e 6b) apresentaram resultados semelhantes e atingiram um valor elevado de poder quando n≥60 para α=5% e n≥95 para α=1%. Com o aumento do número de variáveis de p=2 para p=5, verificou-se que as estatísticas Z_1 e K_1 não apresentaram poder em detectar desvios de normalidade, a não ser com Z_1 em peque nas amostras (Figura 7). Observou-se que com esse aumento do número de variáveis, a estatística K_2 mostrou-se superior em poder à estatística Z_2 (Figura 8). A estatística K_2 apresentou poder elevado quando $n \geq 45$ para α=5% e $n \geq 70$ para α = 1%. A estatística Z_2 apresentou poder elevado quando $n \geq 55$ para

α = 5%, como mostra a Figura 8(a) e n ≥ 80 para α =1% (Figuras 8b). Um tamanho amostral que garante uma boa aproximação para a estatística K_1, nas diferentes distribuições geradas, foi n > 50. Esse tamanho amostral também foi encontrado por Muldholkar et al. (1992), sendo mais adequado para o valor nominal 5%.



(a) (b)

FIGURA 3 – Poder das estatísticas de assimetria para a distribuição bivariada gerada a partir da distribuição de qui-quadrado, 5% (a) e 1% (b).



(a) (b)

FIGURA 4 – Poder das estatísticas de curtose para a distribuição bivariada gerada a partir da distribuição de qui-quadrado, 5% (a) e 1% (b).

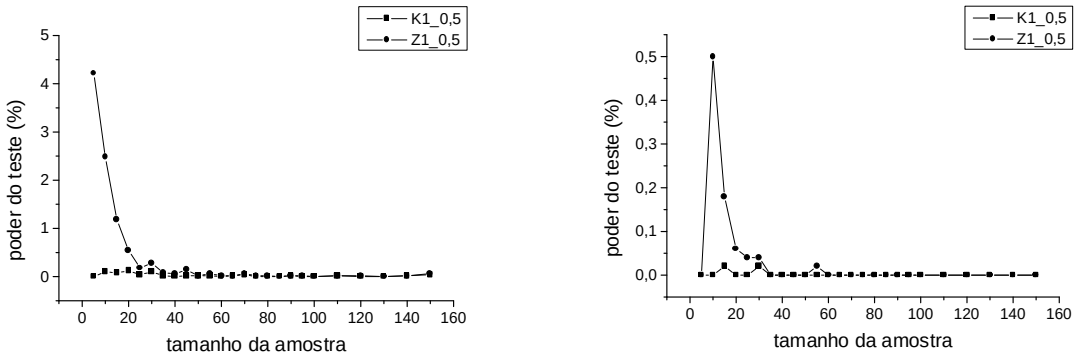


FIGURA 5 – Poder dos testes de assimetria considerando a uniforme bivariada, 5% (a) e 1% (b).

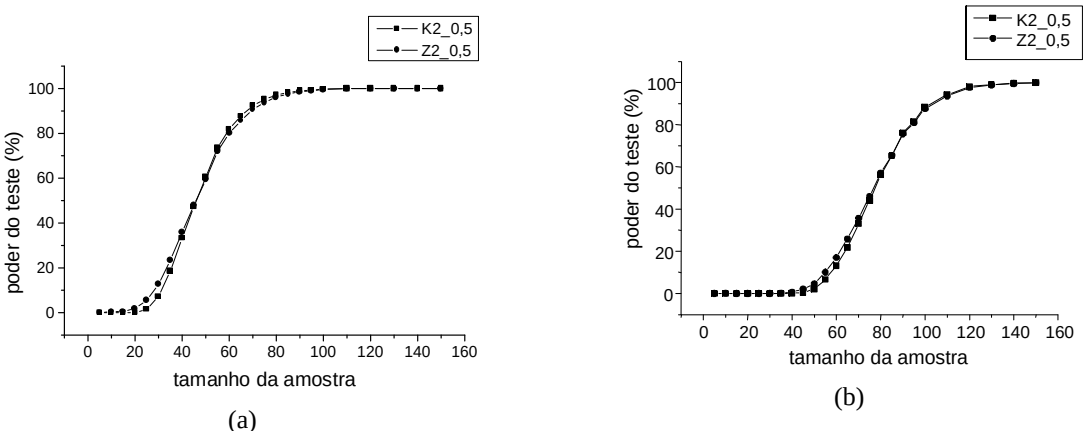


FIGURA 6 – Poder dos testes de curtose considerando a uniforme bivariada, 5% (a) e 1% (b).

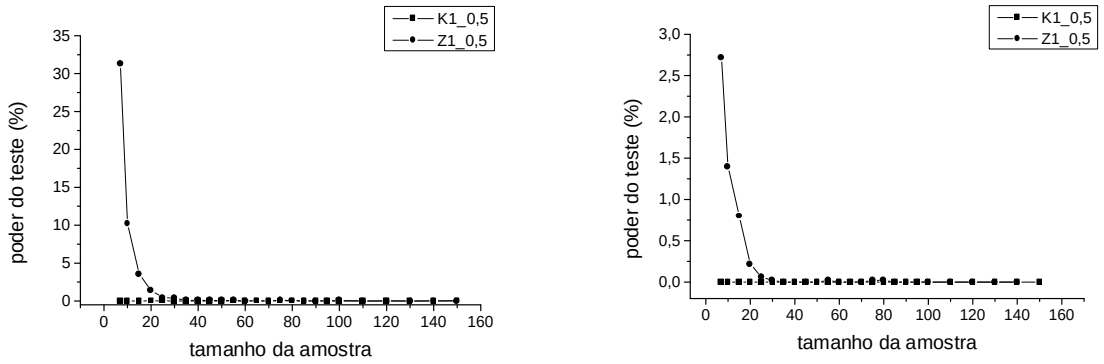


FIGURA 7 – Poder dos testes de assimetria considerando a uniforme pentavariada, 5% (a) e 1% (b).

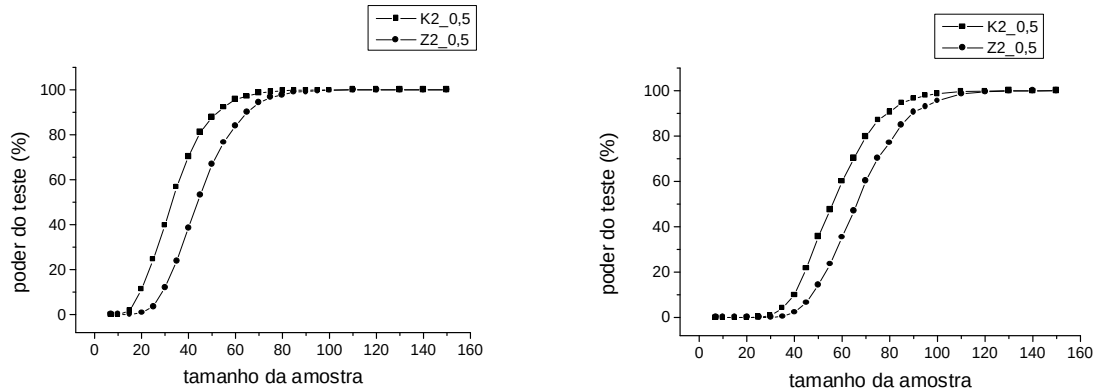


FIGURA 8 – Poder dos testes de curtose considerando a uniforme pentavariada, 5% (a) e 1% (b).

CONCLUSÕES

As diferentes estruturas de correlação não afetam o poder e a taxa de erro tipo I dos testes;

A estatística K_1 é adequada para uso a partir de n≥50 para valores nominais de significância de 5 ou 1%.

A estatística K_2 é assintoticamente adequada para os testes de desvios de curtose para n≥100, independentemente dos valores nominais da significância;

As estatísticas de assimetria, em geral, são mais poderosas do que as de curtose, mas os testes da hipótese nula de normalidade devem considerar tanto os testes de desvios de assimetria como os de curtose conjuntamente.

REFERÊNCIAS BIBLIOGRÁFICAS

BOCK, R. D. Multivariate statistical methods in behavioral research. Chicago: MacGraw–Hill, 1975. 623 p.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. New Jersey: Printice Hall, 1998. 816 p.

MACHADO, S. G. Two statistics for testing multivariate normality. Biometrika, Great Britain, v. 70, n. 3, p. 713-718, Dec. 1983.

MARDIA, K. V. Measures of multivariate skewness and kurtosis with applications. Biometrika, Great Britain, v. 57, n. 3, p. 519-529, Dec. 1970.

MULDHOLKAR, G. S.; McDERMOTT, M.; SRIVASTAVA, D. K. A test p-variate normality. Biometrika, Great Britain, v. 79, n. 4, p. 850-854, Dec. 1992.