Testes robustos baseados na razão de verossimilhanças para independência entre dois grupos de variáveis na presença de *outliers*

Vânia de Fátima Lemes de Miranda [1], **Daniel Furtado Ferreira** [2]

1 Introdução

Atualmente, são coletados dados de vários fenômenos que requer uma análise envolvendo várias variáveis e quanto maior o número delas, mais complexa se torna a interpretação das análises por métodos comuns de estatística univariada.

Uma forma ideal para o estudo de fenômenos complexos é mediante a Estatística Multivariada que, com o avanço das tecnologias computacionais as análises vem se mostrando cada vez mais eficazes, fornecendo maior confiabilidade na tomada de decisões.

A Estatística Multivariada segundo Mingoti (2017), consiste de um conjunto de métodos aplicados em situações onde várias variáveis são medidas simultaneamente em cada elemento amostral. Uma observação multivariada de dimensão p, ou p-variada, é um vetor dado da seguinte forma:

$$\boldsymbol{Y} = [Y_1, Y_2, \cdots, Y_p]^\top,$$

cujas coordenadas Y_1 a Y_p são variáveis aleatórias oriundas de várias medidas de um mesmo elemento amostral, com matriz de observações p-variadas, ou matriz de dados $\boldsymbol{Y}_{n\times p}$. Segundo Ferreira (2011) um conjunto de dados com n medidas em p variáveis podem ser organizados numa matriz, como a seguir

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1k} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2k} & \cdots & Y_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{j1} & Y_{j2} & \cdots & Y_{jk} & \cdots & Y_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nj} & \cdots & Y_{np}. \end{bmatrix}$$

Nesta forma de representação dos dados, cada linha da matriz é um vetor p dimensional de observações multivariadas, e cada coluna, um vetor n-dimensional, das n cópias independentes de uma determinada variável.

Em várias áreas da ciência, como agronomia, econometria, informática, saúde, entre outras é comum obter conjuntos de dados amostrais, resultantes de estudos de vários fenômenos e pesquisas, cujo interesse pode ser avaliar se há independência entre grupos de variáveis, ou seja, na verificação de dependência entre vetores aleatórios. Por exemplo, $\boldsymbol{Y}_{(1)}$ e $\boldsymbol{Y}_{(2)}$ podem representar dois vetores de medições físicas e psicológicas no i-ésimo indivíduo da amostra, com a intenção de verificar se as medições físicas e psicológicas são relacionadas.

A independência entre estes dois grupos de variáveis pode se avaliada pelo teste de razão de verossimilhança (LRT) elaborado por Wilks (1935), que consiste em identificar

[1]UFLA e-mail: *vaniafamat@gmail.com*
[2]UFLA. e-mail: *danielff@des.ufla.br*

se há ou não relação de independência entre os grupos, ou seja, se a covariância entre os dois grupos é nula, isto é, $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Considerando dois grupos de variáveis representados por $\boldsymbol{Y}_{(1)}$ e $\boldsymbol{Y}_{(2)}$ com dimensões $p_1 \times 1$ e $p_2 \times 1$, respectivamente, sendo p_1 e p_2 o número de variáveis em cada grupo, sendo os os vetores de variáveis aleatórias $\boldsymbol{Y}_{(1)}$ e $\boldsymbol{Y}_{(2)}$ normais multivariados com média $\boldsymbol{\mu}$ e covariância $\boldsymbol{\Sigma}$.

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_{(1)} \\ \hline \boldsymbol{\mu}_{(2)} \end{array}\right] \qquad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right],$$

Matriz de covariância amostral $\boldsymbol{S}$ particionada

$$\boldsymbol{S} = \left[\begin{array}{c|c} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \hline \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{array}\right].$$

Para realizar o LRT os vetores $\boldsymbol{Y}_{(1)}$ e $\boldsymbol{Y}_{(2)}$ devem possuir distribuição normal multivariada conjuntamente (JOHNSON; WICHERN, 2007) e o número de variáveis $p = p_1 + p_2$ deve ser menor que o tamanho amostral n.

Este teste de Wilks tem sido discutido e modificado, para os casos em que os vetores não atendem as pressuposições, ou há presença de *outliers*, ou ainda, para um conjunto de dados com alta dimensão.

Muitas vezes, no conjunto de dados há presença de *outliers*, os quais podem comprometer as estimativas obtidas para os parâmetros e fornecer resultados enganosos às inferências. Por isso, antes de aplicar algum método multivariado, deve-se investigar se *outliers* estão presentes nos dados coletados. Segundo Sajesh e Srinivasan (2013) *outliers* são dados inseridos incorretamente ou que não pertencem à população do restante dos dados fornecidos.

Por exemplo, para um vetor aleatório $\boldsymbol{X} \in \mathbb{R}^p$ supostamente normal multivariado, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, considera-se que o vetor de médias amostrais $\bar{\boldsymbol{X}}$ e a matriz de variâncias e covariâncias amostrais $\boldsymbol{S}$ são estimadores que representam bem os componentes do vetor aleatório $\boldsymbol{X}$, possuindo as propriedades de eficiência, consistência e ausência de viés. Porém se for constatado a presença de *outliers* na amostra, estes estimadores são bastante influenciados (SINGH, 1996). Considerando isso, neste trabalho será avaliado o uso do estimador robusto *comedian* para obter as estimativas da matriz de variâncias e covariâncias.

As estatísticas robustas tratam de quantidades para se obter estimativas na presença de *outliers*, seu principal objetivo é atenuar o efeito de *outliers*, bem como preservar a forma, a dispersão e a simetria dos dados reais, preocupando-se com a construção de procedimentos que forneçam resultados confiáveis, em situações nas quais o modelo não esteja em conformidade com os dados.

A teoria da robustez, aliada a métodos Monte-Carlo e métodos numéricos, auxilia no entendimento de problemas de natureza estatística e na rapidez de obtenção de soluções (BUSTOS; YOHAI, 1986).

Para o caso multivariado, são encontrados na literatura vários métodos para obter as estimativas dos parâmetros de locação e escala, os mais utilizados são os métodos baseados no volume mínimo do elipsoide (MVE), métodos baseados no determinante mínimo da matriz de covariâncias (MDC), um método pouco utilizado é o estimador *comedian*.

Uma das vantagens do estimador *comedian* é que ele sempre existe enquanto a covariância requer a existência dos dois primeiros momentos de $\boldsymbol{X}$ e $\boldsymbol{Y}$, e ainda, é simétrico, invariante para a média e covariância, o qual tem forte consistência e normalidade assintótica (FALK, 1997), as estimativas para este estimador podem ser obtidas por meio do pacote *robustbase* do R com a função *covcomed*.

Tendo por base essas informações, o objetivo deste trabalho foi propor três novos testes de independência entre grupos de variáveis utilizando um método robusto na estatística do LRT, aliando-se a eles procedimentos *bootstrap* paramétrico para se obter a distribuição nula das estatísticas dos testes. Além disso, o desempenho dos novos testes e do teste de razão de verossimilhança (LRT) para independência entre dois grupos de variáveis sob distribuições normais e não-normais com a presença de *outliers* será avaliado por meio de simulações Monte Carlo, determinando-se as taxas de erro tipo I e poder.

2 Metodologia

Neste trabalho serão propostos três novos testes para a hipótese H_0 dentre as seguintes hipóteses:

$$H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0} \qquad \text{vs} \qquad H_1 : \boldsymbol{\Sigma}_{12} \neq \mathbf{0}$$

ou

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{array}\right] \qquad \text{vs} \qquad H_1 : \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

O desempenho dos testes propostos e original para independência entre dois grupos de variáveis sob distribuições normais e não-normais com a presença de *outliers* serão realizadas simulações Monte Carlo determinando-se as taxas de erro tipo I e poder.

Para isso, considerou-se uma amostra aleatória de tamanho n de uma distribuição normal multivariada. Inicialmente, aplicou-se o teste original de razão de verossimilhanças, que pressupõe normalidade e cuja estatística possui distribuição assintótica qui-quadrado.

Estatística de Wilks (1935) modificada por Box (1949), dada por

$$\chi^2_{c_1} = -(n-1)(1-C)\left[ln\,|\boldsymbol{S}| - \sum_{i=1}^{2} ln\,|\boldsymbol{S}_{ii}|\right]. \tag{1}$$

em que n é o tamanho da amostra, e ln é o logaritmo neperiano, sob H_0 possui distribuição assintoticamente qui-quadrado com $f = \Gamma_2/2$ graus de liberdade, sendo

$$C = \frac{4\Gamma_3 + 6\Gamma_2}{12(n-1)\Gamma_2} \qquad \text{e} \qquad \Gamma_r = (p_1 + p_2)^r - (p_1^r + p_2^r) \qquad \text{para} \qquad \text{r} = 2,\ 3.$$

Se, a suposição de normalidade multivariada for violada ou ocorrerem *outliers*, este teste tem seu desempenho muito influenciado. Assim, espera-se que os três novos testes propostos são robustos às violações mencionadas. O primeiro deles tem por ideia substituir o estimador $\boldsymbol{S}$ pelo estimador robusto $\boldsymbol{S}^*$ na estatística do teste de razão de verossimilhanças (LTR) (1), que utiliza-se de determinante das matrizes. O segundo teste foi idealizado a partir da utilização da distribuição de *bootstrap* da estatística LRT original (1).

3 Resultados

Nesta seção serão apresentados os novos testes propostos. Essencialmente eles são baseados na estatística LRT de (1) substituindo $\boldsymbol{S}$ por $\boldsymbol{S}^*$, e a distribuição assintótica qui-quadrado pela distribuição *bootstrap*.

3.1 Teste de razão de verossimilhanças robusto (LRTR)

A estatística do teste foi obtida, substituindo-se na LRTO (1) o estimador $\boldsymbol{S}$ pelo estimador robusto $\boldsymbol{S}^*$ obtido pelo método *comedian*.

$$\chi^2_{c_2} = -(n-1)(1-C)\left[ln|\boldsymbol{S}^*| - (ln|\boldsymbol{S}^*_{11}| + ln|\boldsymbol{S}^*_{22}|)\right] \tag{2}$$

em que $\boldsymbol{S}^*$, $\boldsymbol{S}^*_{11}$ e $\boldsymbol{S}^*_{22}$ são os estimadores *comedian* de $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{11}$ e $\boldsymbol{\Sigma}_{22}$, respectivamente. A distribuição da estatística $\chi^2_{c_2}$ sob normalidade e sob H_0 será considerada a χ^2 com $\nu = \Gamma_2/2$ graus de liberdade.

3.2 Teste *bootstrap* paramétrico (LRTB) e (LRTRB)

Como a distribuição da estatística de (1) é apenas assintótica ou aproximadamente assintótica, então buscou-se corrigir o problema usando o método *bootstrap* paramétrico.

Para realizar o LRTB, a partir da amostra original são estimados $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ por $\bar{\boldsymbol{X}}$ e $\boldsymbol{S}$, respectivamente. Para a imposição da hipótese nula, a covariância amostral será modificada por

$$\boldsymbol{S}_0 = \left[\begin{array}{c|c} \boldsymbol{S}_{11} & 0 \\ \hline 0 & \boldsymbol{S}_{22} \end{array}\right].$$

Assim, serão geradas amostras $N_p(\bar{\boldsymbol{X}}, \boldsymbol{S_0})$ de tamanho n, uma vez gerada amostra desta população, a estatística associada deve ser computada por

$$\chi^2_{b1} = -(n-1)(1-C)\left[ln|\boldsymbol{S_b}| - (ln|\boldsymbol{S}^b_{11}| + ln|\boldsymbol{S}^b_{22}|)\right] \tag{3}$$

em que $\boldsymbol{S}_b$ é a matriz de covariâncias obtida na amostra de *bootstrap* gerada.

E para realizar o LRTRB, a partir da amostra original são estimados $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ por $\bar{\boldsymbol{X}}^*$ e $\boldsymbol{S}^*$, respectivamente. Para a imposição da hipótese nula, a covariância amostral será modificada por

$$\boldsymbol{S}^*_o = \left[\begin{array}{c|c} \boldsymbol{S}_{11} & 0 \\ \hline 0 & \boldsymbol{S}_{22} \end{array}\right].$$

Assim, serão geradas amostras $N_p(\bar{\boldsymbol{X}}^*, \boldsymbol{S}^*_0)$ de tamanho n, uma vez gerada amostra desta população, a estatística associada deve ser computada por

$$\chi^2_{b2} = -(n-1)(1-C)\left[ln|\boldsymbol{S}^*_{\boldsymbol{b}}| - (ln|\boldsymbol{S}^{b*}_{11}| + ln|\boldsymbol{S}^{b*}_{22}|)\right] \tag{4}$$

em que $\boldsymbol{S}^*_b$ é a matriz de covariâncias obtida na amostra de *bootstrap* gerada.

Cada processo repetiu B vezes e calculou-se o valor-p comparando com a estatística original, isso será computado na amostra original e em cada amostra de *bootstrap*.

Se χ^2_{bi} for o bi-ésimo valor *bootstrap* paramétrico da estatística então o valor-p teste é dado por

$$Valor - p = \frac{\sum_{b=1}^{B+1} I(\chi^2_{bi} \geq \chi^2_{c_i})}{B+1}, \qquad \text{para } i = 1{,}2 \tag{5}$$

em que $I(A)$ é a função indicadora do evento A e B é o número de reamostragens *bootstrap* para a estatística $\chi^2_{c_i}$ do teste LRT original ou LRTR , com $i = 1$, 2.

Agradecimentos

Agradecimentos a CNPq, CAPES, FAPEMIG, UFLA e UFU.

Referências

BOX, G. E. A general distribution theory for a class of likelihood criteria. *Biometrika*, JSTOR, v. 36, n. 34, p. 317–346, 1949.

BUSTOS, O. H.; YOHAI, V. J. Robust estimates for arma models. *Journal of the American Statistical Association*, Taylor & Francis, v. 81, n. 393, p. 155–168, 1986.

FALK, M. On mad and comedians. *Annals of the Institute of Statistical Mathematics*, Springer, v. 49, n. 4, p. 615–644, 1997.

FERREIRA, D. F. *Estatistica Multivariada.* 2. ed. MG: UFLA, 2011. ISBN 8587692526.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis.* 6. ed. New Jersey, US: Pearson, 2007. v. 1. 808 pages.

MINGOTI, S. A. *Analise de dados atraves de metodos de estatistica multivariada: uma abordagem aplicada.* MG: Editora UFMG, 2017.

SAJESH, T.; SRINIVASAN, M. An overview of multiple outliers in multidimensional data. *Sri Lankan Journal of Applied Statistics*, The Institute of Applied Statistics, Sri Lanka, v. 14, n. 2, p. 87–120, 2013.

SINGH, A. Outliers and robust procedures in some chemometric applications. *Chemometrics and Intelligent Laboratory Systems*, Elsevier BV, v. 33, n. 2, p. 75–100, jun 1996.

WILKS, S. S. On the independence of k sets of normally distributed statistical variables. *Econometrica, Journal of the Econometric Society*, JSTOR, p. 309–326, 1935.