Technological Applied Sciences
ISSN: 1308-7223 (NWSATAS)
ID: 2019.14.4.2A0175

Status : Research Article
Received: 01.08.2019
Accepted: 10.10.2019

Hasan Bulut
Ondokuz Mayıs University, hasan.bulut@omu.edu.tr, Samsun-Turkey

DOI	http://dx.doi.org/10.12739/NWSA.2019.14.4.2A0175
ORCID ID	0000-0002-6924-9651
CORRESPONDING AUTHOR	Hasan Bulut

AN R PACKAGE FOR MULTIVARIATE HYPOTHESIS TESTS: MVTESTS

ABSTRACT

In recent years, the R program is widely used for statistical analysis. Because the R program is open source software, it may contain the R packages developed by users. This study aims to promote an R package entitled MVTests, which consists of multivariate hypothesis tests. By using this package, one performs hypothesis tests, which are used widely and related to each other, such as One-Way MANOVA, multivariate normality tests, Box-M test. In this study, the theoretical background of these tests and their applications with MVTests package are given. Users test easily to their data with MVTests package. Therefore, it is believed that this study is a helpful tool for researchers and users.

Keywords: MVTests, Box-M, MANOVA, Hotelling T Square,
Multivariate Normality Test

1. INTRODUCTION

The multivariate hypothesis tests are generally an adaptation to the univariate hypothesis tests for more variables than one. Like univariate tests, multivariate hypothesis tests are needed using together. For example, the assumptions, which are multivariate normality and homogenous covariance matrices, should satisfy for One Way MANOVA. Therefore, the control of these assumptions is essential before MANOVA is performed. For this reason, we need an only R package which consists of functions to perform these tests. For this aim, we develop an R package, which consists of functions for multivariate hypothesis tests, entitled MVTests (Multivariate Tests) [1]. The MVTests package can perform eight different multivariate hypothesis tests, which are multivariate normality test, Bartlett’s Sphericity test, tests for covariance matrices (single and more groups) and tests for mean vectors (single, two independent samples, multivariate paired and MANOVA). In this study, after we give the theoretical background of multivariate hypothesis tests in this package, we perform applications of these tests with the functions of MVTests package. Moreover, we want to mention that these applications based on version 1.1 of MVTests package. In this reason, it should not be forgotten that MVTests package may consist of additional tests in the future.

2. RESEARCH SIGNIFICANCE

In recent years, the R program is widely used by researchers. In the R program, however, the multivariate tests, which are related to each other, maybe in different packages. This study aims to introduce a new R package, which collects the functions that perform the popular multivariate hypothesis tests. This package is called as MVTests. Moreover, the MVTests package contains multivariate tests, which are not in popular statistical software such as SPSS, SAS, Minitab. Therefore,

How to Cite:
Bulut, H., (2019). An R Package for Multivariate Hypothesis Tests: MVTests,
Technological Applied Sciences (NWSATAS), 14(4):132-138,
DOI:10.12739/NWSA.2019.14.4.2A0175.

it is believed that the MVTests package will contribute to researchers who need performing multivariate tests.

3. MATERIAL AND METHODS

In this section, we promote the theoretical background of eight multivariate tests used in the MVTests package.

3.1. Multivariate Normality Test

In the univariate case, one of the most popular normality tests is the Shapiro-Wilk normality test. Shapiro and Wilk [2] proposed test statistic in (1) for normality test:

$$\mathbf{W_X} = \frac{\tilde{\sigma}_X^2}{s_X^2} \tag{1}$$

where s_X^2 is the sample variance and $\tilde{\sigma}_X^2 = \left[\sum_{i=1}^{n} a_i x_{(i)}\right]^2$. Moreover, $x_{(i)}\ (1,2,\cdots,n)$ rank statistics $\left(x_{(1)} < x_{(2)} < \cdots < x_{(n)}\right)$ and a_i is i^{th} the element of a vector given in (2).

$$a = [a_1 \quad a_2 \quad \dots \quad a_n]' = \frac{\mu_Z' \Sigma_Z^{-1}}{\left(\mu_Z' \Sigma_Z^{-1} \Sigma_Z^{-1} \mu_Z\right)^{0.5}} \tag{2}$$

where $\mu_Z = E(Z)$, $\Sigma_Z = \mathrm{Cov}(Z)$ and Z is is the rank statistics vector of the sample from the standard normal distribution. When $W < k_\alpha$, we reject the null hypothesis, which states that the data has a normal distribution. Villasenor Alva and Estrada [3] generalized this test statistic to the multivariate case. The new test statistics is given in (2).

$$W^* = \frac{1}{p}\sum_{j=1}^{p} W_{Z_j} \tag{3}$$

where, Z_j is the j^{th} element of the Z vector calculated as in (4).

$$Z = S^{1/2}\left(X_j - \overline{X}\right) \tag{4}$$

When $W^* < c_{\alpha;n,p}$, we reject the null hypothesis, which states that the data has a multivariate normal distribution. We need the assumption of multivariate normality in hypothesis tests for mean vectors. For this reason, one should control whether the data has the multivariate normal distribution or not before one performs to test for mean vectors. In the MVTests package, the mvShapiro function performs the multivariate normality test.

3.2. Bartlett's Test for the Covariance Matrix of a Sample

Bartlett's test statistic for the covariance matrix of a sample tests hypothesis as given in (5).

$$\begin{aligned} H_0&: \Sigma = \Sigma_0 \\ H_1&: \Sigma \neq \Sigma_0 \end{aligned} \tag{5}$$

For this aim, Bartlett's test statistic is given in (6).

$$U = (n-1)\left[\ln|\Sigma_0| - \ln|S| + iz\left(S\Sigma_0^{-1}\right) - p\right] \sim \chi^2_{\frac{p(p+1)}{2}} \tag{6}$$

When $U > \chi^2_{\frac{p(p+1)}{2}}$, we reject the null hypothesis. If the sample size is less than 50 $(n < 50)$, we use the test statistic given in (7) instead of in (6) and we reject the null hypothesis when $U' > \chi^2_{\frac{p(p+1)}{2}}$ [4 and 5].

$$U' = \left[1 - \frac{1}{6(n-1)-1}\left(2p + 1 - \frac{2}{p+1}\right)\right] U \sim \chi^2_{\frac{p(p+1)}{2}} \tag{7}$$

In the MVTests package, the Bcov function performs Bartlett's test for a sample covariance matrix.

3.3. Box-M Test for the Covariance Matrix of Multi-group

We need the assumption of homogenous covariance matrices when we perform two samples Hotelling T^2 or MANOVA tests. We use the Box-M test to determine whether the covariance matrices of multi-groups are equal or not. The hypothesis of Box-M test is given in (8):

$$H_0: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$$
$$H_1: \text{At least a } \Sigma_k \text{ is different from other } (k = 1,2,\ldots,g) \tag{8}$$

To test these hypotheses, we sample observations in $n_1, n_2, \ldots, n_g$ size from the multivariate normal distribution, respectively. The covariance matrices of these groups are $S_1, S_2, \ldots, S_g$, respectively. The Box-M test statistic is given in (9).

$$U = -2(1 - c_1)\ln(M) \sim \chi^2_{\frac{p(p+1)(g-1)}{2}} \tag{9}$$

Where;

$$M = \frac{|S_1|^{\frac{n_1-1}{2}} |S_2|^{\frac{n_2-1}{2}} \ldots |S_g|^{\frac{n_g-1}{2}}}{|S_{united}|^{\sum_{i=1}^{g} \frac{n_i-1}{2}}} \quad , S_{united} = \frac{\sum_{i=1}^{g}(n_i-1)S_i}{\sum_{i=1}^{g}(n_i-1)} = \frac{E}{sd_E}$$

$$c_1 = \begin{cases} \left[\frac{2p^2+3p-1}{6(p+1)(g-1)}\right]\left[\sum_{i=1}^{g} \frac{1}{n_i-1} - \frac{1}{\sum_{i=1}^{g}(n_i-1)}\right] & , \quad \text{else} \\ \left[\frac{(g+1)(2p^2+3p-1)}{6g(p+1)(n-1)}\right] & , \quad n_1 = n_2 = \cdots = n_g = n \end{cases}$$

When $U > \chi^2_{\frac{p(p+1)(g-1)}{2};\alpha}$, we reject the null hypothesis [4 and 5]. In MVTests package, BoxM function performs Box-M test.

3.4. One Sample Hotelling T^2 Test

One sample Hotelling T^2 statistic for the mean vector of a sample tests hypothesis as given in (10).

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0 \tag{10}$$

For this aim, one sample Hotelling T^2 statistic is defined as given in (11) [4, 5 and 6].

$$T^2 = n(\overline{X} - \mu_0)' S^{-1} (\overline{X} - \mu_0) \tag{11}$$

This statistic transforms an F statistic as below:

$$\frac{n-p}{p(n-1)} T^2 \sim F_{p;n-p} \tag{12}$$

When $F_h > F_{p;n-p:\alpha}$, we reject H_0 hypothesis. If the H_0 hypothesis is rejected, we use confidence intervals to determine variables affecting the decision to reject. These confidence intervals are calculated as given in (13) [5 and 6].

$$P\left(a\overline{X} - \sqrt{T_t^2 \frac{aSa'}{n}} < a\mu < a\overline{X} + \sqrt{T_t^2 \frac{aSa'}{n}}\right) = 1 - \alpha \tag{13}$$

If the confidence interval involves μ_{0j} for the related parameter, we decide that X_j $(j = 1,2,\ldots,p)$ are not effective on the decision to reject. Otherwise, this parameter is efficient on the decision to reject. In MVTests package, OneSampleHT2 function performs one sample Hotelling T^2 test.

3.5. Hotelling T^2 Test for Two Independent Samples

Two independent samples Hotelling T^2 statistic for the mean vector of two independent samples tests the hypothesis as given in (14).

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2 \tag{14}$$

This test statistics assumes that two groups have multivariate normal distributions $X_1 \sim N_p(\mu_1, \Sigma_1)$, and $X_2 \sim N_p(\mu_2, \Sigma_2)$, respectively. When $\Sigma_1 = \Sigma_2$, which can be tested using BoxM function, we use test statistic as given in (15) [4, 5 and 6].

$$T^2 = \left(\frac{n_1 n_2}{n}\right)(\overline{X}_1 - \overline{X}_2)' S^{-1} (\overline{X}_1 - \overline{X}_2) \sim T^2_{n_1+n_2-2} \tag{15}$$

Where

$$S = \frac{1}{n_1+n_2-2}\left((n_1-1)S_1 + (n_2-1)S_2\right) \quad (16)$$

This statistic transforms an F statistic as below:

$$\frac{n-p-1}{p(n-2)}T^2 = F_{p;n-p-1} \quad (17)$$

When $F_h > F_{p;n-p-1;\alpha}$, we reject the null hypothesis. If the H_0 hypothesis is rejected, we use confidence intervals to determine variables affecting the decision to reject. These confidence intervals are calculated as given in (17) [5 and 6].

$$P\left(a(\overline{X}_1-\overline{X}_2) - \sqrt{T^2_{n_1+n_2-2;\alpha}\frac{n}{n_1n_2}\, aSa'} < a(\mu_1-\mu_2) < a(\overline{X}_1-\overline{X}_2) + \sqrt{T^2_{n_1+n_2-2;\alpha}\frac{n}{n_1n_2}\, aSa'}\right) = 1-\alpha \quad (17)$$

If the confidence interval involves zero for the related parameter, we decide that $X_j\ (j = 1,2,\dots,p)$ are not effective on the decision to reject. Otherwise, this parameter is efficient on the decision to reject. When $\Sigma_1 \neq \Sigma_2$, we use test statistic as given in (18) instead of (15) [5 and 7].

$$T_*^2 = (\overline{X}_1-\overline{X}_2)'S^{-1}(\overline{X}_1-\overline{X}_2) \quad (18)$$

where $S = \frac{1}{n_1}S_1 + \frac{1}{n_2}S_2$. T_*^2 statistic transforms an F statistic as given in (19).

$$T_*^2 \sim \frac{vp}{v-p+1}F_{p;v-p+1} \quad (19)$$

where v is calculated as given in (20).

$$v = \frac{tr\left(\frac{1}{n_1}S_1+\frac{1}{n_2}S_2\right)^2 + \left(tr\left(\frac{1}{n_1}S_1+\frac{1}{n_2}S_2\right)\right)^2}{\frac{1}{n_1-1}\left[tr\left(\frac{1}{n_1}S_1\right)^2 + \left(tr\left(\frac{1}{n_1}S_1\right)\right)^2\right] + \frac{1}{n_1-1}\left[tr\left(\frac{1}{n_1}S_1\right)^2 + \left(tr\left(\frac{1}{n_1}S_1\right)\right)^2\right]} \quad (20)$$

In MVTests package, TwoSamplesHT2 function performs two samples Hotelling T^2 test.

3.4. Multivariate Paired Test

The multivariate paired test statistic for the mean vectors of two dependent samples tests the hypothesis as given in (21).

$$\begin{matrix} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{matrix} \ \text{or} \ \begin{matrix} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{matrix} \quad (21)$$

When $\mu_d = \mu_1 - \mu_2$, equation (21) transforms to equation (10). Therefore, we can use statistic as given in (22).

$$T^2 = n\overline{d}'S_d^{-1}\overline{d} \quad (22)$$

where $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, $S_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i-\overline{d})(d_i-\overline{d})'$ and $d_i = (x_{1i} - x_{2i})$. This statistic transforms into the F statistic as below.

$$\frac{n-p}{p(n-1)}T^2 \sim F_{p;n-p} \quad (23)$$

When $F_h > F_{p;n-p:\alpha}$, we reject H_0 hypothesis [4]. In MVTests package, Mpaired function performs the multivariate paired test.

3.5. One Way Multivariate Analysis of Variance (MANOVA)

One Way MANOVA (Multivariate Analysis of Variance) tests whether the mean vectors of groups, which are the least 3, are equal to each other or not. In MANOVA, we assume that the groups have the multivariate normal distribution and homogenous covariance matrix. The hypotheses are given as following.

$$\begin{matrix} H_0: \mu_1 = \mu_2 = \dots = \mu_g \\ H_1: \text{At least a } \mu_k \text{ is different from other} \end{matrix} \quad (24)$$

We may use three different statistics to test these hypotheses. These statistics base on matrices as below:

$$T = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(X_{ik}-\overline{X})(X_{ik}-\overline{X})'$$

$$\mathrm{B} = \sum_{k=1}^{g}\sum_{i=1}^{n_k} n_k(\bar{X}_{.k} - \bar{X})(\bar{X}_{.k} - \bar{X})'$$

$$\mathrm{W} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(X_{ik} - \bar{X}_{.k})(X_{ik} - \bar{X}_{.k})'$$

- **Wilks' Lambda Statistic:** Wilks' Λ statistic is defined given as (25) and it transforms into an F statistics as in (26).

$$\Lambda = \frac{|W|}{|B+W|} = \frac{|W|}{|T|}, 0 \le \Lambda \le \tag{25}$$

$$F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\frac{df_2}{df_1} \sim F_{df_1,df_2} \tag{26}$$

where $V_E = n - g, V_H = g - 1, df_1 = pV_H, df_2 = wt - \frac{1}{2}(pV_H - 2)$

$$t = \sqrt{\frac{p^2V_H^2-4}{p^2+V_H^2-5}}, w = V_E + V_H - \frac{1}{2}(p + V_H + 1)$$

When $F_h > F_{df_1,df_2;\alpha}$ I we reject the null hypothesis.

- **Roy's Largest Root:** Roy's largest root test statistic use eigenvalues of the $W^{-1}B$ matrix. Let's be $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ of $W^{-1}B$ matrix's eigenvalues. Roy's largest root statistic is defined as given in (27).

$$\theta = \frac{\lambda_1}{1+\lambda_1} \tag{27}$$

By using θ statistic, we can obtain an F statistic as follows.

$$F = \frac{\lambda_1(V_E-r+V_H)}{r} \sim F_{r,(V_E-r+V_H)} \tag{28}$$

where $r = \max(p, V_H)$. When $F_h > F_{r,(V_E-r+V_H);\alpha}$, we reject the null hypothesis.

- **Hotelling-Lawley Trace Test:** Hotelling-Lawley Trace test statistic is defined in (29).

$$HL = nT_0^2 \sim \chi^2_{p(g-1)} \tag{29}$$

where T_0^2 is the trace of $W^{-1}B$ matrix. Moreover, T_0^2 can transform into an F statistic as given in (30).

$$F = \frac{2(s\breve{N}+1)T_0^2}{s^2(2m+s+1)} \sim F_{s(2m+s+1),2(s\breve{N}+1)} \tag{30}$$

where $\breve{N} = \frac{(V_E-p-1)}{2}, m = \frac{|V_H-p|-1}{2}, s = \min(V_H, p)$. When $F_h > F_{s(2m+s+1),2(s\breve{N}+1);\alpha}$, we reject the null hypothesis. When the null hypothesis is rejected using these test statistics, we decide that the mean vectors of groups are different from each other. Moreover, we benefit from confidence intervals to determine the source of these differences. These confidence intervals are calculated as follows.

$$P\left[\sum_{k=1}^{g} c_k\, a\, \hat{\varphi}_k - \sqrt{\lambda_\alpha^*\, aWa' \sum_{k=1}^{g}\frac{c_k^2}{n_k}} \le \sum_{k=1}^{g} c_k\, a\varphi_k \le \sum_{k=1}^{g} c_k\, a\varphi\, \hat{\varphi}_k + \sqrt{\lambda_\alpha^*\, aWa' \sum_{k=1}^{g}\frac{c_k^2}{n_k}}\right] = 1 - \alpha$$
(31)

where c_k values is contrast constants, **a** is a vector, which consists of zero values except j^{th} element. The j^{th} element of this vector is 1. The critical table value is also calculated as follows.

$$\lambda_\alpha^* = \frac{(V_E-r+V_H)}{r\left(F_{r,(V_E-r+V_H);\alpha}\right)} \tag{32}$$

When any confidence interval contains zero, the means of related groups are not different with regards to j^{th} variable. When the assumption of the homogenous covariance matrix is not satisfied, James' test statistic, which is given in (33), should be used [9].

$$J = \sum_{i=1}^{g}(\boldsymbol{X}_i - \boldsymbol{X})'\boldsymbol{W}_i(\boldsymbol{X}_i - \boldsymbol{X}) \tag{33}$$

where $\boldsymbol{W}_i = \left(\frac{1}{n_i}\boldsymbol{S}_i\right)^{-1}$, $\boldsymbol{S}_i$ is the covariance matrix of i^{th} group and $\overline{\boldsymbol{X}} = \left(\sum_{i=1}^{g}\boldsymbol{W}_i\right)^{-1}\sum_{i=1}^{g}\boldsymbol{W}_i\overline{\boldsymbol{X}}_i$. James' statistic is compared with the critical value of χ^2 distribution [9 and 10].

$$2h(\alpha) = \chi_r^2(A + B\chi_r^2) \qquad (34)$$

where; $r = p(g-1), A = 1 + \frac{1}{2r}\sum_{i=1}^{g}\frac{\left[tr\left(\boldsymbol{I}_p - \boldsymbol{W}^{-1}\boldsymbol{W}_i\right)\right]^2}{n_i - 1}$ and

$$B = \frac{1}{r(r+2)}\sum_{i=1}^{g}\left\{\frac{tr\left[\left(\boldsymbol{I}_p - \boldsymbol{W}^{-1}\boldsymbol{W}_i\right)\right]^2}{n_i - 1} + \frac{\left[tr\left(\boldsymbol{I}_p - \boldsymbol{W}^{-1}\boldsymbol{W}_i\right)\right]^2}{2(n_i - 1)}\right\}$$

In MVTests package, Manova function performs the One Way MANOVA.

- **Sphericity Test:** Bartlett's sphericity test statistic tests the hypothesis as follows.

$$H_0: R = I$$
$$H_1: R \neq I \qquad (35)$$

The test statistic is denoted as:

$$\chi^2 = -\left[(n-1) - \frac{1}{6}(2p+5)\right]\log(|R|) \sim \chi^2_{\frac{p(p-1)}{2}} \qquad (36)$$

When $\chi^2 > \chi^2_{\frac{p(p-1)}{2}}$, we reject the null hypothesis. When the null hypothesis is rejected, we can use the dimension reduction methods such as principal component analysis or factor analysis. In MVTests package, Bsper function performs Bartlett's sphericity test.

4. CONCLUSION

In this study, we have introduced multivariate hypothesis tests, which are widely used and related to each other. Then, the applications of these tests have been given by basing on functions, which are in the MVTests package that can perform these tests. Moreover, the MVTests package can perform the different tests based on approaches such as mvShapiro, James's MANOVA, Nel and Merve's approach to two independent samples Hotelling $\mathbf{T}^2$. These different tests are not in the popular package programs such as SPSS, Minitab, SAS. Therefore, this study supplies fast, free and efficient software to researchers, who want to perform multivariate hypothesis tests. From this aspect, we believe that this study can be used in scientific studies and contribute to science.

REFERENCES

[1] Bulut, H., (2018). MVTests: Multivariate Hypothesis Tests. R package version 1.1. URL: https://cran.r-project.org/web/packages/MVTests/MVTests.pdf

[2] Shapiro, S.S. and Wilk, M.B., (1965). An Analysis of Variance Test for Normality (complete samples). Biometrika, 52(3/4):591-611.

[3] Villasenor Alva, J.A. and Estrada, E.G., (2009). A Generalization of Shapiro-Wilk's Test for Multivariate Normality. Communications in Statistics—Theory and Methods, 38(11):1870-1883.

[4] Rencher, A.C., (2003). Methods of Multivariate Analysis (Vol. 492). John Wiley & Sons.

[5] Bulut, H., (2018). R Uygulamaları ile Çok Değişkenli İstatistiksel Yöntemler. Nobel Akademik Yayıncılık, Ankara, Türkiye.

[6] Tatlidil, H., (1996). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Cem Web., Ankara, Türkiye.

[7] Nel, D.G. and Van der Merwe, C.A., (1986). A Solution to the Multivariate Behrens-Fisher Problem. Communications in

Statistics-Theory and Methods, 15(12):3719-3735.

[8] https://cran.r-project.org/

[9] James, G.S., (1954). Tests of Linear Hypotheses in Univariate and Multivariate Analysis When the Ratios of the Population Variances are Unknown. Biometrika, 41(1/2):19-43.

[10] Tsagris, M.T., (2014). Multivariate Statistical Functions in R. Athens, Nottingham and Abu Halifa (Kuwait).