Colloquium Biometricum 43
2013, 29–35

HOW TO GET P-VALUE FOR MULTIVARIATE NORMALITY

Zofia Hanusz, Joanna Tarasińska

Department of Applied Mathematics and Computer Science
University of Life Sciences
Akademicka 13, 20-950 Lublin, Poland
e-mail: joanna.tarasinska@up.lublin.pl

Summary

The idea how to get *p*-value for multivariate normality using R software is presented. This idea combines *p*-values of Shapiro-Wilk *W* statistics, calculated for principal components of the sample covariance matrix, with four statistics based on combination of independent *p*-values. The example with R commands is given.

Keywords and phrases: multivariate normality, Shapiro Wilk *W* statistic, R software, *p*-value

Classification AMS 2010: 62G10

1. Introduction

Normality is the basic assumption in the analysis of both univariate and multivariate data.

In univariate case the Shapiro-Wilk *W* test statistic (Shapiro and Wilk 1965, 1968) is considered as a very powerful one, especially for small sample size. In its classical form the tables with proper coefficients and critical values are necessary. However, Royston (1992) gave the way of normalization of *W* statistic. His transformation allows to get *p*-value for normality and it is implemented in statistical software. There are a lot of attempts to adapt *W* statistic for multivariate case. Some of such ideas are given in Malkovich and

Afifi (1973), Royston (1983), Fattorini (1986), Srivastava and Hui (1987), Mudholkar et al. (1995), Liang et al. (2009), Villasenor and Estrada (2009).

In this paper we propose to join the Srivastava and Hui's idea (Srivastava and Hui, 1987) of taking the principal components (PC) of the sample covariance matrix with the idea of combinations of p-values for independent tests (Zwet and Oosterhoff 1967, Mudholkar et al. 1995).

The idea is illustrated by an example with sample size $n = 28$ and $k = 4$ variates.

2. Shapiro–Wilk *W* statistic and its Royston's approximation

Shapiro–Wilk W statistic (Shapiro and Wilk, 1965) for testing univariate normality is defined by:

$$W = \frac{\left(\sum_{j=1}^{n} a_j x_{(j)}\right)^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2},$$

where $x_{(1)} \le \cdots \le x_{(n)}$ are ordered statistics, $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ and a_j are elements of the vector $\mathbf{a}' = \frac{\mathbf{m}'\mathbf{V}^{-1}}{\left(\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m}\right)^{\frac{1}{2}}}$ where $\mathbf{m} = E\left[x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right]'$ and $\mathbf{V} = \left[\operatorname{cov}(x_{(i)}, x_{(j)}\right]$ are expected value and covariance matrix of ordered statistics, respectively. Small values of W indicate nonnormality.

Royston (1992) gave the idea of normalization of W statistic so that the p-value of the test can be calculated as upper tail of the standard normal distribution. This normalization is as follows.

For $4 \le n \le 11$, the transformed variable

$$w = -\ln[-2.273 + 0.459n - \ln(1 - W)]$$

has got, under normality, normal distribution with parameters

$$\mu = 0.5440 - 0.39978n + 0.025054n^2 - 0.0006714n^3,$$

$$\sigma = \exp\left(1.3822 - 0.77857n + 0.062767n^2 - 0.0020322n^3\right).$$

For $12 \le n \le 2000$, the variable

$$w = \ln(1 - W)$$

has got, under normality, normal distribution with parameters

$$\mu = -1.5861 - 0.31082x - 0.083751x^2 + 0.0038915x^3,$$

$$\sigma = exp\left(-0.4803 - 0.082676x + 0.0030302x^2\right),$$

where $x = \ln n$.
This idea is implemented in the procedure 'shapiro.test' in R environment.

3. Combination of independent *p*-values

Let $p_1, p_2, \ldots, p_k$ denote the *p*–values for *k* independent tests of hypothesis $H_0^1, H_0^2, \ldots, H_0^k$, respectively. Now, let us consider the null hypothesis $H_0 = \left\{H_0^1, H_0^2, \ldots, H_0^k\right\}$. The following statistics can be considered as test statistics for H_0 (Zwet and Oosterhoff 1967, Mudholkar et al. 1995):

$$W_F = -2\sum_{i=1}^{k} \ln p_i ,$$

$$W_L = A^{-\frac{1}{2}} \sum_{i=1}^{k} \ln\left(\frac{p_i}{1 - p_i}\right) \text{ with } A = \frac{\pi^2 k(5k+2)}{15k+12} ,$$

$$W_N = \sum_{i=1}^{k} \Phi^{-1}(1 - p_i) ,$$

$$W_T = \min(p_i),$$

where *F*, *L*, *N* and *T* refer to Fisher, logit, Liptak and Tippet combination method. Under H_0 the statistics are distributed as follows:

$$W_F \sim \chi^2_{2k},$$

$$W_L \overset{as}{\sim} t_{5k+4},$$

$$W_N \overset{as}{\sim} N(0,k).$$

W_T is distributed as the minimum of *k* uniform variates.

Lower tails of W_L, W_T and upper tails of W_F, W_N indicate nonnormality.

4. *P* – values for multivariate normality

Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be an iid sample from a *k*-variate population. We are interested in testing the null hypothesis:

$$H_0 : (\mathbf{X}_1,\ldots,\mathbf{X}_n) \text{ is a sample from normal distribution } N_k(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown.

Let us consider the principal components of sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$ where $\overline{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j$ and let p_i $(i = 1,2,\ldots k)$ be the *p*-value of Shapiro-Wilk *W* statistic for the *i*-th principal component. The principal components, under H_0, are asymptotically independent and normally distributed. Thus we can combine $p_i's$ to obtain statistics W_F W_L, W_N and W_T given in section 3.

Thus, our proposition of getting *p*–values for multivariate normality is as follows.

1. For *k*-variate sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ find the principal components (PC's) of sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})' .$$

2. Use 'shapiro.test' in R to get *p*-values for each of PC.
3. Combine these *p*-values to get one of the statistics W_L, W_T, W_F, W_N (or all of them).
4. Calculate *p*-value for multivariate normality as $1-F_{\chi^2_{2k}}(W_F)$, $F_{t_{5k+1}}(W_L)$, $1-F_{N(0,k)}(W_N)$, $1-(1-W_T)^k$, where F denotes the suitable cumulative distribution function.

5. Example

Let us illustrate the tests by the example from Rao (1948) and recalled by Srivastava (2002).

The data consists of weights in centigrams of cork borings for the north (N), east (E), south (S) and west (W) sides of the trunks for 28 trees.

The matrix of observations is

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	45	48	54	57	43

Let 'data' be the $n \times k$ matrix with data. The R commands are below:

```
> S=cov(data)
> H=eigen(S)$vectors
> Y=data%*%H
> k=4
> p_values=array(,c(k))
> for (i in(1:k)){p_values[i]=shapiro.test(Y[,i])$p.value}
> WF=-2*sum(log(p_values))
> pvalueWF=1-pchisq(WF,2*k,ncp=0)
> print(pvalueWF)
[1] 0.008357556
> A=pi^2*k*(5*k+2)/(15*k+12)
> WL=A^(-0.5)*sum(log(p_values))
> pvalueWL=pt(WL,5*k+1,ncp=0)
> print(pvalueWL)
[1] 0.003716385
> WN=sum(qnorm(1-p_values,0,1))
> pvalueWN=1-pnorm(WN,0,sqrt(k))
> print(pvalueWN)
[1] 0.01109188
> WT=min(p_values)
> pvalueWT=1-(1-WT)^k
> print(pvalueWT)
[1] 0.03795216
```

Thus *p*-values are as follows: 0.00836 for W_F , 0.00372 for W_L , 0.01109 for W_N and 0.03795 for W_T and each of the four tests rejects multivariate normality. For some other results of testing normality in this example see also Hanusz and Tarasinska (2006).

References

Fattorini L. (1986). Remarks on the use of the Shapiro-Wilk statistic for testing multivariate normality. *Statistica* 46, 209–217.

Hanusz Z., Tarasinska J. (2006). Certain measure in graphical methods for checking multivariate normality. *Colloquium Biometryczne* 36, 149–158.

Liang J., Tang M.L., Chan P.S. (2009). A generalized Shapiro-Wilk W statistic for testing high-dimensional normality. *Computational Statistics and Data Analysis* 53, 3883–3891.

Malkovich J.F., Afifi A.A. (1973). On tests for multivariate normality. *JASA* 68, 176–179.

Mudholkar G.S., Srivastava D.K., Lin C.T. (1995). Some *p*-variate adaptations of the Shapiro-Wilk test of normality. *Commun. Statist.-Theory Meth*. 24 (4), 953–985.

Rao C.R. (1948). Tests of significance in multivariate analysis. *Biometrika* 35, 58–79.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0, URL http://www.R–project.org.

Royston J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Appl. Statist*. 32, No. 2, 121–133.

Royston J. P. (1992). Approximating the Shapiro-Wilk *W*-test for non-normality. *Statistics and Computing* 2, 117–119.

Shapiro S.S., Wilk M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611.

Shapiro S.S., Wilk M.B. (1968). Approximations for the null distribution of the W statistic. *Technometrics* 10, No. 4 , 861–868.

Srivastava M.S., Hui T.K. (1987). On assessing multivariate normality based on Shapiro-Wilk W statistic. *Statist. Prob. Lett*. 5, No. 1, 15–18.

Srivastava M.S. (2002). *Methods of multivariate statistics*. J. Wiley & Sons, New York.

Villasenor Alva J.A., Estrada E.G. (2009). Generalization of Shapiro-Wilk's test for multivariate normality. *Commun. Statist.-Theory Meth*. 38 (11), 1870–1883.

Zwet W.R., Oosterhoff J. (1967). On the combination of independent test statistics. *AMS* 38, No. 3, 659–680.