Colloquium Biometricum 38
2008, 87–93

REMARKS ON APPROXIMATED TESTS BASED ON SHAPIRO-WILK'S STATISTIC

Zofia Hanusz, Joanna Tarasińska

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

Summary

Some remarks concerning Johnson's S_B transformation of critical values for Shapiro-Wilk's W statistic and its goodness of fit to standard normal quantiles for different sample sizes are presented. Two another tests for multivariate normality based on Johnson's S_B transformation of W statistic are proposed.

Keywords and phrases: test for normality, test for multivariate normality, Shapiro-Wilk statistic

Classification AMS 2000: 62H15, 62G10

1. Introduction

There are a lot of tests for checking normality of data but as the best one, especially for small sample size, is considered Shapiro-Wilk's test. Let us remind this test.

Let $x_1, x_2, \dots, x_n$ be independent identical distributed variables from normal distribution and $x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$ be the ordered values of $x_1, x_2, \dots, x_n$. Shapiro-Wilk's W statistic (Shapiro and Wilk, 1965) is given 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 a_j are constants tabulated by Shapiro and Wilk (1965) and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Small values of W indicate nonnormality and critical values W_α for $n = 3(1)50$ are also tabulated in Shapiro and Wilk (1965).

Shapiro and Wilk propose also test based on the following transformation, using Johnson's (1949) S_B distribution

$$G(W) = \gamma + \delta \ln\left(\frac{W - \varepsilon}{1 - W}\right) , \qquad (1.1)$$

where $G(W)$ is distributed as standard normal. Tables for γ, δ and ε for sample sizes $n = 3(1)50$ are given in Shapiro and Wilk (1968) and Srivastava (2002). The lower tail of normal distribution indicates nonnormality.

This Shapiro-Wilk's approximated test can be easily adopted to multivariate case. For example Srivastava and Hui (1987) propose the following approach. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be independently distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ be the sample mean and sample covariance matrix, respectively. Let $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_p)$ be an orthogonal matrix such that $\mathbf{S} = \mathbf{H}\mathbf{D}_u\mathbf{H}'$, where $\mathbf{D}_u = diag(u_1, \dots, u_p)$ and $u_1 \ge \dots \ge u_p$ are eigenvalues of $\mathbf{S}$. Define $y_{ij} = \mathbf{h}_i \mathbf{x}_j$, $i = 1, \dots, p$; $j = 1, \dots, n$. Now let us take p univariate Shapiro-Wilk's statistics

$$W(i) = \frac{1}{nu_i}\left[\sum_{j=1}^{n} a_j y_{i(j)}\right]^2 \quad i = 1, \dots, p$$

where $y_{i(1)} \leq \ldots \leq y_{i(n)}$ are ordered statistics for i-th principal component. Srivastava and Hui (1987) propose to take as a test statistic for multivariate normality the following one:

$$M = -2\sum_{i=1}^{p} \ln\left[\Phi\left(G\left(W\left(i\right)\right)\right)\right],$$

where $G\left(W\left(i\right)\right)$ is the transformation (1.1) and $\Phi\left(\cdot\right)$ is the cumulative distribution function of standard normal distribution. It must be mentioned that Srivastava and Hui (1987) took unbiased estimate of $\mathbf{\Sigma}$, $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)\left(\mathbf{x}_i - \overline{\mathbf{x}}\right)'$, instead of maximum likelihood one. However it was pointed by Hanusz and Tarasińska (2008) as a mistake. The statistic M, under multivariate normality of $\mathbf{x}_i$, is approximately distributed as χ^2_{2p} as $W\left(i\right)$ are asymptotically independent and $\Phi\left(G\left(W\left(i\right)\right)\right)$ are uniformly distributed. Thus, large values of M indicate nonnormality. Of course the statistic M can be also used in univariate case $p = 1$.
Srivastava and Hui (1987) also propose

$$M_2 = \min_{i=1,\ldots,p}\left\{W\left(i\right)\right\}$$

as the second test statistic, the distribution of which is approximately given by

$$\Pr\left(M_2 \leq x\right) = 1 - \left[1 - \Phi\left(G\left(x\right)\right)\right]^p .$$

This test rejects normality for small values of M_2.

2. Remark on the approximation (1.1)

Statisticians using the approximation (1.1) should not suppose that normality of $G\left(W\right)$ improves when sample size increases. For each n the value ε is

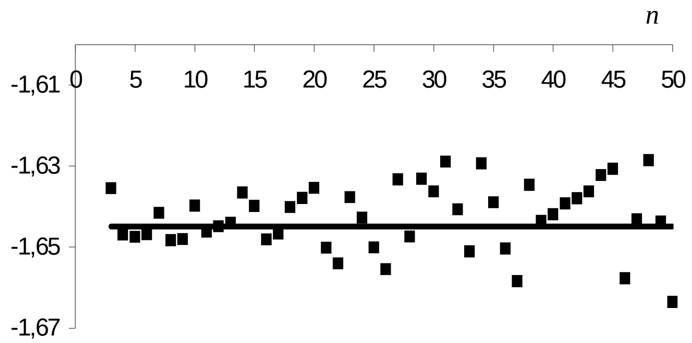

Fig. 1. Values of G_α and quantile of standard normal distribution for $\alpha = 0.05$

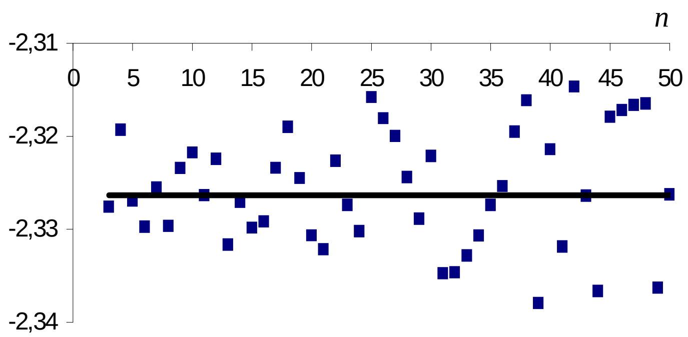

Fig. 2. Values of G_α and quantile of standard normal distribution for $\alpha = 0.01$

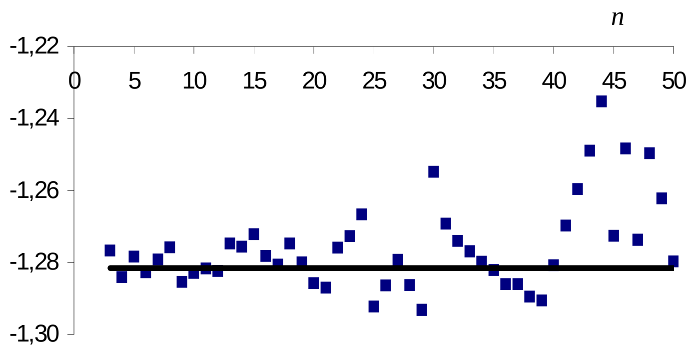

Fig. 3. Values of G_α and quantile of standard normal distribution for $\alpha = 0.1$

taken as the minimum of *W*, namely $\frac{na_1^2}{n-1}$. The constants γ and δ were evaluated by Shapiro and Wilk (1968) by means of simple least squares regression $u=\frac{1}{\delta}z-\frac{\gamma}{\delta}$, where $u=\ln\frac{W-\varepsilon}{1-W}$ and *z* is distributed as standard normal. Values employed in regression were quantiles z_p and W_p for $p=0.01, 0.02, 0.05(0.05)0.25(0.25)0.75(0.05)0.95, 0.98, 0.99$. Such values of *p* were to weight the goodness of fit in the tails.

Thus the transformation of critical values W_α via (1.1) needn't fit better the proper quantiles of normal distribution for larger *n*. It can be even worse. The figures 1, 2 and 3 give values of $G_\alpha=\gamma+\delta\ln\left(\frac{W_\alpha-\varepsilon}{1-W_\alpha}\right)$ as a function of sample size ($n=3,\ldots,50$) for $\alpha=0.05$, 0.01 and 0.1, it means for usually used significance levels in testing normality. Solid lines in the figures point $\Phi^{-1}(\alpha)$.

3. Preliminary suggestion of two other tests for multivariate normality

Following Srivastava and Hui's (1987) idea of Shapiro-Wilk's (1965,1968) statistics for principal components of covariance matrix **S** and using transformation (1.1) for them two another tests for multivariate normality can be proposed.

The first of them can be based on test function $\sqrt{p}\,\overline{V}$, where $\overline{V}$ is the average of $V_i=G(W(i))$. Under normality $\sqrt{p}\,\overline{V}$ has got asymptotic standard normal distribution. The lower tail of it indicates nonnormality.

The second test statistic can be taken as $V=\min_{i=1,\ldots,p}\{V_i\}$. Its asymptotic distribution, under normality, is given by the cumulative distribution function of the form $F(x)=1-[1-\Phi(x)]^p$. The lower tail of this distribution indicates nonnormality. Thus critical values can be found as $V_\alpha=\Phi^{-1}\left(1-\sqrt[p]{1-\alpha}\right)$.

Both proposed tests require investigations as to their type I error and power but preliminary researches seem to be promising. Figures 4 and 5 present histograms of proposed test functions obtained for 1000 samples of size 10, *p*=2 generated from normal distribution. Solid lines in the figures show theoretical probability density function. In spite of the fact that proposed test statistics are not affine invariant, the histograms look similar for correlation coefficient $\rho = 0$ and ± 0.9.

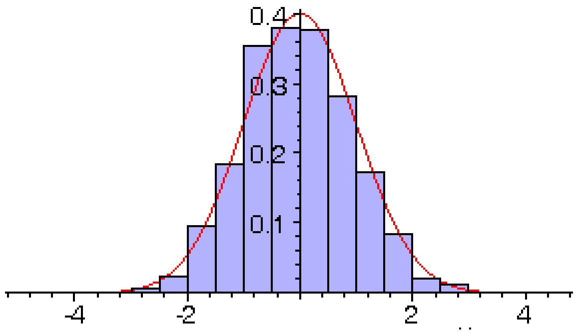

Fig. 4. Histogram for test function $\sqrt{p}\,\overline{V}$ for 1000 samples generated from normal distribution

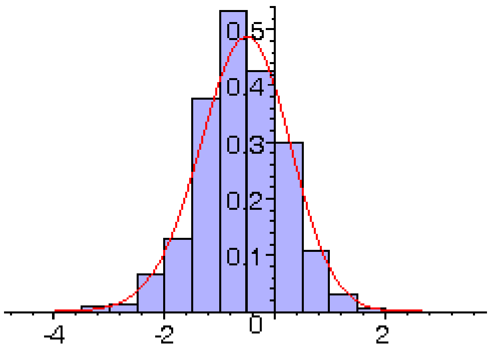

Fig. 5. Histogram for test function *V* for 1000 samples generated from normal distribution

Acknowledgments

This research was supported by MSHE grant N N310 3280 33.

References

Hanusz Z., Tarasińska J. (2008). A note on Srivastava and Hui's tests of multivariate normality. (to publish) *Journal of Multivariate Analysis*.

Johnson N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* 36, 149-176.

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, 861-866.

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

Srivastava M.S., Hui T.K. (1987). On assessing multivariate normality based on Shapiro-Wilk W statistic. *Statistics & Probability Letters* 5, 15-18.

UWAGI O PRZYBLIŻONYCH TESTACH OPARTYCH NA STATYSTYCE SHAPIRO-WILKA

Streszczenie

Podane są pewne spostrzeżenia dotyczące przekształcenia wartości krytycznych statystyki Shapiro-Wilka poprzez transformację związaną z systemem rozkładów S_B Johnsona oraz ich dobroci dopasowania do odpowiednich kwantyli rozkładu normalnego w zależności od liczebności próby. Zaproponowane są także dwa nowe testy wielowymiarowej normalności oparte o transformacje statystyki Shapiro-Wilka.

Słowa kluczowe: test normalności, test wielowymiarowej normalności, sytatystyka Shapiro-Wilka

Klasyfikacja AMS 2000: 62H15, 62G10