Colloquium Biometricum 44
2014, 35−41

NORMALIZATION OF SHAPIRO−WILK TEST WITH KNOWN MEAN

Zofia Hanusz, Joanna Tarasińska

Department of Applied Mathematics and Computer Science
University of Life Sciences, Głęboka 28, 20-612 Lublin, Poland
e-mails: zofia.hanusz@up.lublin.pl, joanna.tarasinska@up.lublin.pl

Summary

The paper concerns the adaptation W_0 of the Shapiro-Wilk W statistic to the case of testing normality with known mean (Hanusz et al., 2012) and gives the way for normalization of the W_0 statistic using Johnsons (1949) S_B transformation. Thus the p-values of W_0 can be easily computed.

Keywords and phrases: Shapiro-Wilk *W* test, normality, known mean

Classification AMS 2010: 62F03

1. Introduction

Normality is one of the most common assumptions when we use statistical procedures. There are a lot of tests for checking normality (for the review, see for example Thode, 2002). One of the mostly known and applied tests is the Shapiro-Wilk W test (Shapiro and Wilk, 1965), based on statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2},$$

where $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ are the ordered values of the sample $(X_1, X_2, \ldots, X_n)$ and a_i are tabulated coefficients. Small values of W indicate nonnormality. In literature, this test is recommended as very powerful (Thode, 2002; Razali and Wah, 2011) for the null hypothesis that a random variable X is normally distributed with unknown parameters.

Hanusz et al. (2012) gave adaptation of this test to the case of known mean, i.e. to the case when the null hypothesis is of the form:

H_0: X is normally distributed with a known mean μ_0. (1.1)

This modification of the Shapiro-Wilk W statistic is of the following form:

$$W_0 = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \mu_0\right)^2}. \tag{1.2}$$

The hypothesis (1.1) is rejected at significance level α if W_0 is less than the critical value $W_0(\alpha;n)$. The critical values of W_0 for different sample sizes and $\alpha = 0.1, 0.05, 0.01$ were given in Hanusz et al. (2012). However, it would be more convenient to have a transformation of W_0 with a known null distribution.

The aim of this paper is to use Johnson's (1949) S_B distribution in order to normalize W_0. The normalization is made in the same way as it is given by Shapiro and Wilk (1968) for the W statistic.

2. Normalization for the null distribution of W_0 statistic

The Johnson's S_B (1949) distribution can be used to get normal approximation of a bounded test statistic T, where

$$Z = \gamma + \delta \ln\left(\frac{T-\varepsilon}{\lambda - T}\right)$$

is approximately distributed as standard normal. The parameters ε and λ are the minimum and maximum attainable values of statistic *T*, respectively. The values of γ and δ may be evaluated by Monte Carlo study.

Let us describe this approximation after Shapiro and Wilk (1968). In the case of the Shapiro-Wilk *W* statistic we have $\lambda = 1$ and $\varepsilon = \frac{na_1}{n-1}$ for all sample sizes *n*. The normalizing coefficients γ and δ were found by Shapiro and Wilk (1968) in the following way. For different sample sizes *n* they made the simple least squares regression of the empirical sampling values of

$$u(p) = \ln \frac{W(p) - \varepsilon}{1 - W(p)}$$

on the *p*-th quantile z_p of the standard normal distribution, where $W(p)$ denoted the *p*-th empirical sampling quantile of *W*. The regression leads to estimates of $-\gamma/\delta$ and $1/\delta$ from which γ and δ may be obtained. Shapiro and Wilk employed the following values of *p*:

$$p = 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 0.8, 0.85, 0.9, 0.95, 0.98, 0.99$$

and gave the tables for γ, δ and ε. The lower tail of statistic $Z = \gamma + \delta \ln\left(\frac{W-\varepsilon}{1-T}\right)$ indicates nonnormality.

The same method may be used for the W_0 statistic. In this case we have $\lambda = 1$ and $\varepsilon = 0$ as the denominator $\sum_{i=1}^{n}(X_i - \mu_0)^2$ in (1.2) can be arbitrarily large.

In our study, the least squares regression of $\ln \frac{W_0(p)}{1 - W_0(p)}$ on z_p was based on 1,000,000 pseudorandom samples of size from 3 to 50, generated from standard normal distribution. The values of γ and δ, such that

$$Z = \gamma + \delta \ln \frac{W_0}{1 - W_0} \tag{2.1}$$

has approximately standard normal distribution, are listed in Table 1. The lower tail of statistic (2.1) indicates that the hypothesis (1.1) should be rejected.

Table 1. The normalizing constants for W_0 for sample sizes n

n	γ	δ	n	γ	δ
3	-0.3137	0.5551	27	-3.9346	1.4606
4	-0.6479	0.7282	28	-4.0077	1.4703
5	-0.9586	0.851	29	-4.0770	1.4783
6	-1.2299	0.9384	30	-4.1538	1.4891
7	-1.4778	1.0092	31	-4.2084	1.4935
8	-1.6950	1.0671	32	-4.2782	1.503
9	-1.8960	1.1157	33	-4.3354	1.5086
10	-2.0790	1.1573	34	-4.4017	1.5172
11	-2.2470	1.1929	35	-4.4593	1.5241
12	-2.4039	1.2238	36	-4.5088	1.5272
13	-2.5513	1.2517	37	-4.5621	1.5336
14	-2.6821	1.2755	38	-4.6152	1.5382
15	-2.8104	1.2979	39	-4.6749	1.5467
16	-2.9320	1.3181	40	-4.7186	1.5495
17	-3.0400	1.335	41	-4.7771	1.5574
18	-3.1553	1.3542	42	-4.8195	1.5597
19	-3.2563	1.3698	43	-4.8711	1.5659
20	-3.3584	1.3847	44	-4.9137	1.5693
21	-3.4511	1.3983	45	-4.9706	1.5769
22	-3.5365	1.4095	46	-5.0118	1.5797
23	-3.6320	1.4236	47	-5.0512	1.5826
24	-3.7067	1.4319	48	-5.0908	1.5858
25	-3.7869	1.4431	49	-5.1470	1.5935
26	-3.8624	1.452	50	-5.1795	1.5954

After plotting the values (n, γ) and (n, δ), we can see that there exist functions $\gamma(n)$ and $\delta(n)$ which describe regression of γ and δ on sample size n with R^2 near to one.

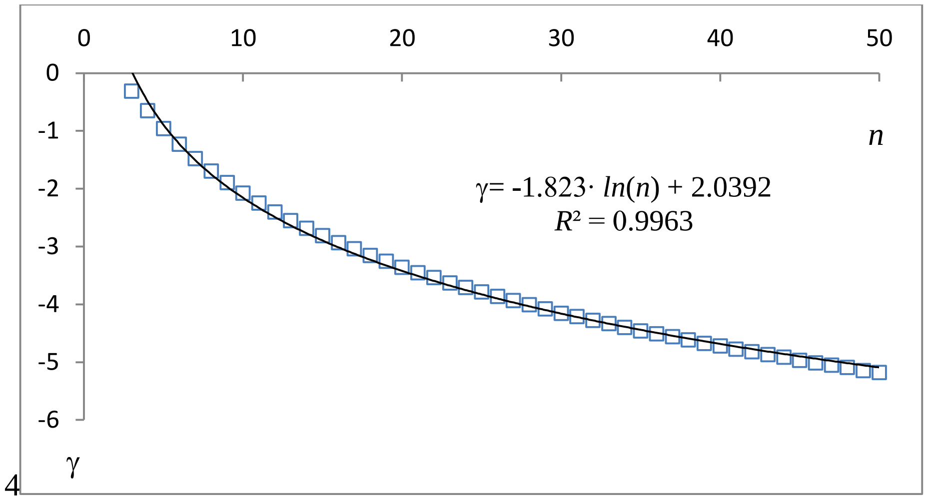

Fig. 1. The scatter plot and regression line $\gamma(n)$

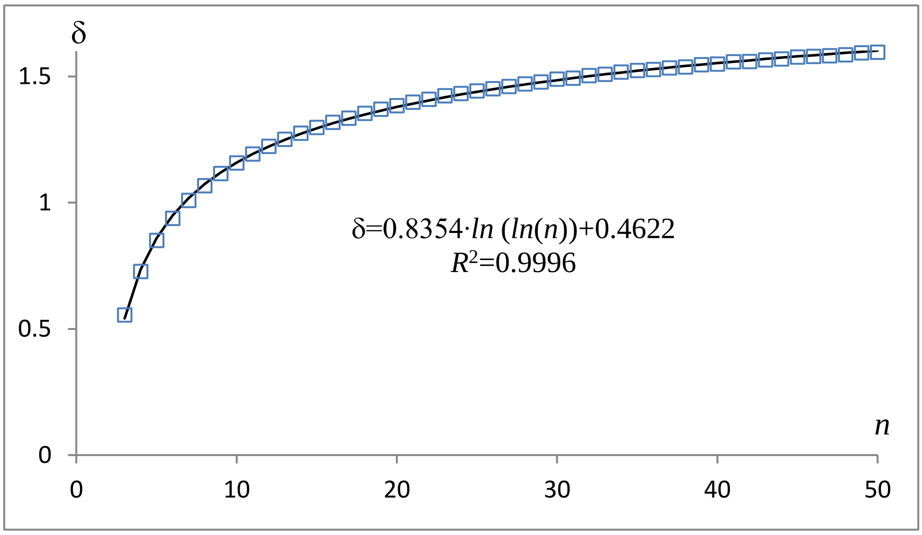

Fig. 2. The scatter plot and regression line $\delta(n)$

For the values from Table 1 we have

$$\gamma = -1.823\ln(n) + 2.0392 \tag{2.2}$$

with $R^2 \approx 0.9963$ (see Fig.1) and

$$\delta = 0.8354\ln(\ln(n)) + 0.4622 \tag{2.3}$$

with $R^2 \approx 0.9996$ (see Fig.2).

The *p*-value for statistic W_0 can be found as $\Phi\left(\gamma + \delta\ln\frac{W_0}{1-W_0}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution, γ and δ are given by formulas (2.2) and (2.3), respectively.

3. Illustration

To illustrate the use of results presented in Section2, let us consider the data consisted of weights in centigrams of cork borings for the north and south sides of the trunks for 28 trees (Srivastava, 2002, ex. 1.2.1). Let us assume we are interested in verifying the null hypothesis that the difference *D* between weights for north and south sides is normally distributed with mean zero:

$$H_0 : D \sim N(0, \sigma^2)$$

The value of statistic W_0 can be determined as $W_0 = W \cdot \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n} X_i^2}$ where the value of *W* may be got for example by “shapiro.test” in R program. For our data we get $W_0 = 0.9273209$. The critical value for statistic W_0 at significance level 0.05 is $W_0(0.05;28) = 0.8287$ (Hanusz et al., 2012). Thus the null hypothesis is not rejected.

However, following the results in Section 2 we do not need table with critical values for W_0. It is sufficient to use formulas (2.2) and (2.3) to get

$$\gamma = -1.823\ln(28) + 2.0392 \approx -4.03541$$

$$\delta = 0.8354\ln(\ln(28)) + 0.4622 \approx 1.467716.$$

Now, we are able to compute the p-value for the test:

$$\Phi\left(\gamma+\delta\ln\frac{W_0}{1-W_0}\right)\approx\Phi(-4.03541+1.467716\cdot\ln(12.75913))\approx$$
$$\approx\Phi(-0.29824)\approx 0.383.$$

If the Shapiro-Wilk W test for normality of data and then classical t-test for hypothesis $H_0:\mu_D=0$ are applied, we get p-values 0.1009 for W test and 0.574 for t- test. In our opinion the test based on W_0, generating only one p-value, is more useful.

4. Conclusion

For testing null hypothesis about normality with known mean the test based on normalizing transformation of statistic W_0, i.e. the test based on $Z=\gamma+\delta\ln\frac{W_0}{1-W_0}$, may be used. There is no need for tables with coefficients γ and δ for different sample sizes n, as there are well-fitting regression lines $\gamma(n)$ and $\delta(n)$ given by (2.2) and (2.3). The test gives possibility to obtain p-value which is the lower tail of standard normal distribution.

References

Hanusz Z., Tarasinska J., Zieliński W. (2012). Adaptation of Shapiro-Wilk test to the case of known mean. *Colloquium Biometricum* 42, 43−50.

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

Razali N.M., Wah Y.B. (2011). Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modelling and Analytics*, Vo.2, No 1, 21−33.

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.

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, New York.

Thode H.C. (2002). *Testing for normality*. Marcel Dekker Inc.