Statistics and Computing (1992) **2**, 117–119

Approximating the Shapiro–Wilk W-test for non-normality

PATRICK ROYSTON

Department of Medical Physics, Royal Postgraduate Medical School, Ducane Road, London W12 0NN

Received October 1991 and accepted October 1991

A new approximation for the coefficients required to calculate the Shapiro–Wilk W-test is derived. It is easy to calculate and applies for any sample size greater than 3. A normalizing transformation for the W statistic is given, enabling its P-value to be computed simply. The distribution of the new approximation to W agrees well with published critical points which use exact coefficients.

Keywords: Non-normality, Shapiro–Wilk W-test

1. Introduction

Shapiro and Wilk's (1965) W-test is a well-established and powerful test of departure from normality. Royston (1982a) extended the test from the original sample size of $n = 50$ to $n = 2000$. The extension was based on Shapiro and Wilk's own suggested approximation to the coefficients used in the calculation of W, but no evaluation was made of its accuracy. Unfortunately, the use of exact rather than approximate coefficients requires matrix manipulations which are both computationally demanding and memory-hungry. In this paper, a new approximation to the coefficients is produced which is simple, convenient and satisfactory for any sample size. A normalizing transformation of the W statistic is provided to enable simple calculation of the P-value of the test for $4 \le n \le 2000$, though extrapolation beyond 2000 appears justified.

2. New approximation

Suppose $y_1 < y_2 < \ldots < y_n$ is an ordered sample of size n to be tested for non-normality. The W statistic is defined by

$$W = \left(\sum a_i y_i\right)^2 \Big/ \sum (y_i - \bar{y})^2$$

where $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ is such that $(n-1)^{-\frac{1}{2}} \Sigma\, a_i y_i$ is the best linear unbiased estimate (BLUE) of the standard deviation of the y_i, assuming normality. The exact value of $\mathbf{a}$ is

$$\mathbf{a} = (\mathbf{m}^{\mathrm{T}} V^{-1} V^{-1} \mathbf{m})^{-\frac{1}{2}} \mathbf{m}^{\mathrm{T}} V^{-1}$$

where V is the covariance matrix of the order statistics of a sample of n standard normal random variables with expectation vector $\mathbf{m}$. The vector $\mathbf{a}$ is antisymmetric, that is, $a_n = -a_1$ and, for odd n, $a_{[n/2]+1} = 0$. Also $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$, so for $n = 3$, $a_n = -a_1 = 2^{-\frac{1}{2}}$. The values of $\mathbf{m}$ and V may be calculated using the algorithms of Royston (1982b) and Davies and Stephens (1978), respectively, supplemented by that of Shea and Scallan (1988) for V_{11}, the variance of the largest order statistic. Since $\mathbf{V}$ is a symmetric $n \times n$ matrix, storage of at least $n + n(n+1)/2$ reals followed by $n \times n$ matrix inversion is needed to compute $\mathbf{a}$. For large samples (say, $n > 250$), the task is beyond most statistical packages and desktop machines, even with today's technology.

As noted indirectly by Verrill and Johnson (1988), the W-test is asymptotically equivalent to the Shapiro and Francia (1972) statistic

$$W' = \left(\sum b_i y_i\right)^2 \Big/ \sum (y_i - \bar{y})^2$$

where $b_i = (\mathbf{m}^{\mathrm{T}}\mathbf{m})^{-\frac{1}{2}} m_i$, and to the Weisberg and Bingham (1975) statistic $(\Sigma\, c_i y_i)^2 / \Sigma\, (y_i - \bar{y})^2$, where $c_i = (\tilde{\mathbf{m}}^{\mathrm{T}}\tilde{\mathbf{m}})^{-\frac{1}{2}} \tilde{m}_i$, $\tilde{m}_i = \Phi^{-1}\{(i - \frac{3}{8})/(n + \frac{1}{4})\}$ and Φ is the normal cdf. Thus $\mathbf{b}$ and $\mathbf{c}$ may be regarded as first approximations to $\mathbf{a}$, $\mathbf{c}$ being particularly simple to compute since it only requires the inverse normal cdf.

0960-3174/92 $03.00 +.12

In fact, up to a constant, $\mathbf{b}$ and $\mathbf{c}$ differ from $\mathbf{a}$ mainly in the first two components (also in the last two, but $a_i = -a_{n-i+1}$). Because of its convenience, $\mathbf{c}$ was chosen as the basis for approximating $\mathbf{a}$. Polynomial (quintic) regression analysis of $a_n - c_n$ and $a_{n-1} - c_{n-1}$ on $x = n^{-\frac{1}{2}}$ for $4 \le n \le 1000$ gave the following equations:

$$\tilde{a}_n = c_n + 0.221157x - 0.147981x^2 - 2.071190x^3 + 4.434685x^4 - 2.706056x^5$$

$$\tilde{a}_{n-1} = c_{n-1} + 0.042981x - 0.293762x^2 - 1.752461x^3 + 5.682633x^4 - 3.582663x^5$$

Then normalizing the remaining $\tilde{m}_i$ by writing

$$\phi = (\tilde{\mathbf{m}}^{\mathrm{T}}\tilde{\mathbf{m}} - 2\tilde{m}_n^2)/(1 - 2\tilde{a}_n^2) \quad \text{if } n \le 5$$

$$= (\tilde{\mathbf{m}}^{\mathrm{T}}\tilde{\mathbf{m}} - 2\tilde{m}_n^2 - 2\tilde{m}_{n-1}^2)/(1 - 2\tilde{a}_n^2 - 2\tilde{a}_{n-1}^2) \quad \text{if } n > 5$$

we have

$$\tilde{a}_i = \phi^{-\frac{1}{2}}\tilde{m}_i$$

for $i = 2, \ldots, n-1$ ($n \le 5$) or $i = 3, \ldots, n-2$ ($n > 5$).

3. Some comparisons

A comparison between various approximations to a_n is given in Table 1. Shapiro and Wilk's (1965) approximation for $n > 20$ (Royston, 1982a, for $n > 50$) is inadequate; even their 'exact' values for $n \le 20$ are incorrect. The new approximation, $\tilde{a}_n$, is accurate to ± 1 in the fourth decimal place. The asymptotic approximation $c_n = (\tilde{\mathbf{m}}^{\mathrm{T}}\tilde{\mathbf{m}})^{-\frac{1}{2}}\tilde{m}_n$ converges slowly to a_n but is still about 6% too low at $n = 1000$. The maximum absolute errors over all i ($1 \le i \le n$) and selected n ($4 \le n \le 1000$, as in Table 1) are 0.0689 for Shapiro and Wilk, 0.0010 for $\tilde{\mathbf{a}}$ and 0.0335 for c_n. The error in $\tilde{\mathbf{a}}$ is greatest in the fourth component ($i = 4$).

Critical percentage points of the W statistic using different approximations to $\mathbf{a}$ are shown in Table 2. The points were obtained by simulation of pseudorandom normally distributed samples with 10 000 repetitions per sample size. The values for exact $\mathbf{a}$ are taken from Table 1 of Verrill and Johnson (1988) and were based on 5000 repetitions. The points using the $\tilde{\mathbf{a}}$ approximation agree closely with those of Verrill and Johnson (the variation is greater at the 0.01 critical point, as would be expected). Royston (1982a) gives much lower critical values for $n > 80$, because the important a_n/a_{n-1} coefficient ratio is far too large, which places inappropriate weight on the extreme sample values relative to the next most extreme. The 0.05 critical points for the $\tilde{\mathbf{a}}$ approximation with $n > 2000$ were obtained by extrapolation using the normalization of W described in the next section. Note how closely they agree with Verrill and Johnson's exact 0.05 points calculated from asymptotic theory.

Table 1. *Exact values of a_n and some approximations*

n	Exact a_n	Shapiro–Wilk	$\tilde{a}_n$	c_n
4	0.6873	0.6869	0.6873	0.6800
5	0.6645	0.6647	0.6646	0.6516
6	0.6430	0.6431	0.6430	0.6258
7	0.6232	0.6258	0.6231	0.6029
8	0.6051	0.6039	0.6051	0.5826
9	0.5887	0.5854	0.5887	0.5645
10	0.5737	0.5694	0.5737	0.5483
12	0.5474	0.5429	0.5474	0.5204
15	0.5149	0.5124	0.5150	0.4868
20	0.4733	0.4758	0.4734	0.4449
25	0.4417	0.4450	0.4418	0.4138
30	0.4166	0.4255	0.4167	0.3895
40	0.3786	0.3964	0.3786	0.3530
50	0.3506	0.3751	0.3506	0.3265
75	0.3034	0.3392	0.3034	0.2823
100	0.2729	0.3158	0.2728	0.2539
125	0.2510	0.2988	0.2509	0.2335
150	0.2341	0.2855	0.2340	0.2179
200	0.2094	0.2657	0.2093	0.1951
250	0.1918	0.2514	0.1918	0.1788
350	0.1677	0.2311	0.1677	0.1566
500	0.1451	0.2114	0.1452	0.1357
750	0.1228	0.1910	0.1229	0.1151
1000	0.1089	0.1778	0.1091	0.1023

4. Normalization of W statistic

For $n = 3$, the cdf of W is $(6/\pi)\{\sin^{-1}(W^{\frac{1}{2}}) - \sin^{-1}(0.75^{\frac{1}{2}})\}$ (Shapiro and Wilk, 1965), small values indicating non-normality. For $4 \le n \le 11$, a three-parameter lognormal distribution was found to fit the empirical (simulation) distribution of $\ln(1 - W)$ adequately. A two-parameter lognormal fitted the upper half of the distribution of $1 - W$ for $n > 11$. The mean, μ, and standard deviation, σ, of w (transformed W) were estimated as the median and (95th centile − median)/1.6449 of w, respectively, and were smoothed as follows. For $4 \le n \le 11$:

$$w = -\ln[\gamma - \ln(1 - W)]$$

$$\gamma = -2.273 + 0.459n$$

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

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

For $12 \le n \le 2000$:

$$x = \ln n$$

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

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

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

Table 2. *Some critical points for W*

n	Royston (1982a)		Based on $\tilde{a}$		Verrill and Johnson (1988)	
	0.05	0.01	0.05	0.01	0.05	0.01
20	0.904	0.868	0.904	0.866	0.905	0.855
30	0.928	0.902	0.930	0.903	0.929	0.893
40	0.941	0.922	0.945	0.924	0.945	0.917
60	0.9538	0.9411	0.9605	0.9459	0.9606	0.9442
80	0.9602	0.9505	0.9691	0.9579	0.9694	0.9565
100	0.9642	0.9561	0.9746	0.9654	0.9747	0.9652
150	0.9698	0.9640	0.9822	0.9760	0.9821	0.9750
250	0.9748	0.9710	0.9888	0.9850	0.9890	0.9853
500	0.97945	0.97707	0.99411	0.99218	0.99419	0.99200
1000	0.98305	0.98159	0.99692	0.99594	0.99683*	–
2000	0.98602	0.98514	0.99839	0.99790	0.99837*	–
4000	–	–	0.99915	0.99890	0.99916*	–
5000	–	–	0.99931	0.99910	0.99932*	–

*Asymptotic points

To find the P-value for W, $z = (w - \mu)/\sigma$ is referred to the upper tail of $N(0, 1)$.

Worked example. The following is a random sample of size 10 from a lognormal distribution: 48.4, 49.0, 59.5, 59.6, 60.7, 88.8, 98.2, 109.4, 169.1, 227.1. The first five approximate expected order statistics ($\tilde{\mathbf{m}}$) for $n = 10$ are -1.5466, -1.0005, -0.6554, -0.3755, -0.1226 and the corresponding approximate coefficients ($\tilde{\mathbf{a}}$) are -0.5737, -0.3290, -0.2143, -0.1228, -0.0401. Thus $W = 158.594^2/31136 = 0.8078$. For $n = 10$, we have $\gamma = 2.317$, $\mu = -1.6198$, $\sigma = 0.11544$, giving $w = -1.3779$, $z = (w - \mu)/\sigma = 2.095$ and $P = 0.018$. The data depart significantly from normality at the 0.02 significance level. The value of W using exact coefficients ($\mathbf{a}$) is also 0.8078 to four decimal places, so the approximation here introduces negligible error.

5. Comments

The W-test may be modified for use in different situations, including testing normality in censored samples (Verrill and Johnson, 1988), multivariate normality (Royston, 1983) and goodness-of-fit of the three-parameter lognormal distribution (Royston, 1991a). Its power characteristics are well known and may be summarized by saying that it is strongest against short-tailed (platykurtic) and skew distributions and weakest against symmetric moderately long-tailed (leptokurtic) distributions. A negative feature is its sensitivity to the presence of ties in the data, caused by rounding or imprecise measurement. It then rejects the null hypothesis too often, but Royston (1991b) shows how to obtain conservative P-values under such circumstances. In practical terms, a normal probability plot should always accompany the W-test to provide qualitative information about the shape of the sample distribution.

References

Davies, C. S. and Stephens, M. A. (1978) Algorithm AS128. Approximating the covariance matrix of normal order statistics. *Applied Statistics*, **27**, 206–212.

Royston, J. P. (1982a) An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, **31**, 115–124.

Royston, J. P. (1982b) Algorithm AS177. Expected normal order statistics (exact and approximate). *Applied Statistics*, **31**, 161–165.

Royston, J. P. (1983) Some techniques for assessing multivariate normality based on the Shapiro–Wilk W. *Applied Statistics*, **32**, 121–133.

Royston, P. (1991a) Estimation, reference ranges and goodness-of-fit for the three-parameter lognormal distribution. *Statistics in Medicine*, in press.

Royston, P. (1991b) Aspects of goodness-of-fit tests for normality. I. Use of the tests when the data contain ties. *Technometrics*, submitted.

Shapiro, S. S. and Francia, R. S. (1972) An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, **67**, 215–216.

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

Shea, B. L. and Scallan, A. J. (1988) AS R72. A remark on Algorithm AS128: approximating the covariance matrix of normal order statistics. *Applied Statistics*, **37**, 151–155.

Verrill, S. and Johnson, A. (1988) Tables and large-sample distribution theory for censored-data correlation statistics for testing normality. *Journal of the American Statistical Association*, **83**, 1192–1197.

Weisberg, S. and Bingham, C. (1975) An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics*, **17**, 133–134.