The Statistician (1993) 42, pp. 37–43

A toolkit for testing for non-normality in complete and censored samples

PATRICK ROYSTON

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

Abstract. The Shapiro–Wilk W and Shapiro–Francia W' statistics are convenient and powerful tests of departure from normality. Modifications are described to allow the use of W and W' with grouped and with singly censored data, and of W with log-normally distributed data. Methods are given to enable the P value of each test to be calculated under these different circumstances. It is hoped that software developers will thereby be encouraged to incorporate one or other of the tests in their statistical products.

1 Introduction

Since their introduction (Shapiro & Wilk, 1965; Shapiro & Francia, 1972) the Shapiro–Wilk W and Shapiro–Francia W' tests of departure of a sample from a hypothesized normal distribution have proved both statistically powerful and popular among users. Their popularity may be inferred from their adoption by the developers of several major statistical packages. Part of the appeal probably lies in their interpretation as measures of linearity of the normal probability plot, particularly W', which is the squared Pearson correlation coefficient between the ordered sample values and the expected normal order statistics.

In practice, several 'irrelevant' types of departure from normality may occur, departures which disturb the null distributions of the W and W' test statistics but which do not necessarily imply non-normality of the underlying random variable (or of a transformation of it). The three main instances are grouping of data due to rounding or imprecise measurement, censoring and estimation of an extra parameter, e.g. in the shifted log-normal distribution. The first two produce an inflated Type I error rate (too many 'significant' tests) while the third has the opposite effect, resulting in a 'supernormal' distribution and corresponding lack of power.

The purpose of the present paper is to describe recent modifications to the W and W' tests which overcome most of the aforementioned problems, providing the user with a 'toolkit' of algorithms which are easy to program. The method for evaluating the P values of the tests with censored data has not appeared before. A secondary aim is to demonstrate the usefulness of a simple general approach to transforming the null distribution of a test statistic to normality.

2 The tests

2.1 Complete samples

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

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

where $\mathbf{a}=(a_1,\ldots,a_n)'$ is a vector of weights that is antisymmetric (i.e. $a_n=-a_1$ and, for odd n, $a_{[n/2]+1}=0$) and normalized so that $\mathbf{a}'\mathbf{a}=1$. For the W test, $(n-1)^{-1/2}\Sigma a_i y_i$ is the BLUE (best linear unbiased estimate) of the SD of the y_i, assuming normality. The exact value of $\mathbf{a}$ is

$$\mathbf{a}=(\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{-1/2}\mathbf{m}'\mathbf{V}^{-1}$$

where $\mathbf{V}$ is the covariance matrix of the order statistics of a sample of n standard normal random variables with expectation vector $\mathbf{m}=(m_1,\ldots,m_n)'$. For the W' test,

$$\mathbf{a}=(\mathbf{m}'\mathbf{m})^{-1/2}\mathbf{m}$$

so the weights are essentially the normal order statistics $m_1,\ldots,m_n$. The asymptotic null distributions of $\mathbf{W}$ and $\mathbf{W}'$ are identical, but convergence is very slow (Verrill & Johnson, 1988).

2.2 *Calculation of weights* (*a*)

The values of $\mathbf{m}$ and $\mathbf{V}$ may be calculated using the algorithms of Royston (1982a) and Davies & Stephens (1978), respectively, supplemented by that of Shea & 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 $\mathbf{a}$. Unfortunately, the task is computationally demanding and is beyond the reach of most statistical packages. Some versions of certain packages, such as GLIM and BMDP, do allow the user to link in externally written routines to perform the required matrix calculations.

In practice, the values of $\mathbf{m}$ may be approximated by the Blom scores $\tilde{\mathbf{m}}$ where

$$\tilde{m}_i=\Phi^{-1}\{(i-3/8)/(n+1/4)\} \qquad (i=1,\ldots,n)$$

and Φ^{-1} is the inverse normal cumulative distribution function. Weights $\mathbf{c}=(\tilde{\mathbf{m}}'\tilde{\mathbf{m}})^{-1/2}\tilde{\mathbf{m}}$ may be used to calculate W'. Royston (1992a) provided the following approximation $\tilde{\mathbf{a}}$ to the weights for the W test. It is based on $\tilde{\mathbf{m}}$ and some simple polynomials in $u=n^{-1/2}$ and removes the need for any matrix computations. For $n\geqslant 4$, we define

$$\tilde{a}_n=c_n+0{\cdot}221157u-0{\cdot}147981u^2-2{\cdot}071190u^3+4{\cdot}434685u^4-2{\cdot}706056u^5$$

$$\tilde{a}_{n-1}=c_{n-1}+0{\cdot}042981u-0{\cdot}293762u^2-1{\cdot}752461u^3+5{\cdot}682633u^4-3{\cdot}582633u^5$$

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

$$\begin{aligned}\phi&=(\tilde{\mathbf{m}}'\tilde{\mathbf{m}}-2\tilde{m}_n^2)/(1-2\tilde{a}_n^2) && \text{if } n\leqslant 5\\ &=(\tilde{\mathbf{m}}'\tilde{\mathbf{m}}-2\tilde{m}_n^2-2\tilde{m}_{n-1}^2)/(1-2\tilde{a}_n^2-2\tilde{a}_{n-1}^2) && \text{if } n>5\end{aligned}$$

we have

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

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

2.3 *Singly censored samples*

It is readily seen that W and W' are the squared product–moment correlations between $\mathbf{y}$ and $\mathbf{a}$, allowing their extension to censored samples. I consider only single (without loss of generality, upper-tail) censoring. Suppose that a proportion Δ of the sample of size n is censored (Type II censoring) and that $k=[n(1-\Delta)]$. Then values $\mathbf{y}_\Delta=(y_1,\ldots,y_k)'$ remain and a normal plot of $\mathbf{y}_\Delta$ against $\mathbf{a}_\Delta$ may still be made. The W and W' statistics are the squared correlations between $\mathbf{y}_\Delta$ and $\mathbf{a}_\Delta$.

2.4 Grouped data

Let $r_1, \ldots, r_n$ be the ranks of the observations, using average ranks for tied y-values in the usual manner. Modified Blom scores may be calculated as $\tilde{m}_i = \Phi^{-1}\{(r_i - 3/8)/(n + 1/4)\}$, together with corresponding approximate weights $\tilde{\mathbf{a}}$. W and W' are then calculated in correlation-coefficient form as described in Section 2.3. The resulting tests are slightly conservative, but without the modification the Type I error rates are grossly inflated (Royston, 1992 b).

2.5 Three-parameter log-normal distribution (3PL)

In the 3PL, $\ln(\mathbf{y} - \gamma)$ is normally distributed. If the value of γ is known or hypothesized, no modification to the tests is required. However, if γ is estimated from the sample that is subsequently tested for non-normality, an adjustment is required (Royston, 1992 c).

3 Transformation of the null distributions of *W* and *W'* to normality

3.1 Standard case

The following approach applies quite generally to a test statistic X having a continuous distribution, though for clarity it is described in terms of W and W'. The aim is to find a transformation g such that $g(X)$ has an (approximately) normal distribution with mean μ and variance σ^2. It is convenient to define g such that large positive values of $g(X)$ correspond to significant (small) values of W and W'. The P value for X is then $1 - \Phi^{-1}(Z)$ where $Z = (g(X) - \mu)/\sigma$. In practice the distribution of X is unknown and is determined by simulation over a set of selected values of n, typically in the range $5 \leqslant n \leqslant 2000$. The transformation is found by trial-and-error, though the Box–Cox $g(X) = (X^\lambda - 1)/\lambda$ and/or the shifted logarithmic $g(X) = \ln(X - \gamma)$ functions are often satisfactory. The parameters μ, σ, λ and γ, or simple transformations of them, are then smoothed as functions (usually polynomials) of $u = f(n)$; typically $u = \ln(n)$.

Table 1 gives the details of $g(X)$ and of associated parameter smoothing functions for W (Royston, 1992a) and W' (Royston, 1992d). The coefficients in Table 1 are values of $\beta_0, \ldots, \beta_k$ for polynomials $\Sigma_{i=0}^{k} \beta_i u^i$ in u.

3.2 Modifications

The distribution of Z may still be approximately normal in non-standard situations, but not necessarily with mean $\mu_Z = 0$ or SD $\sigma_Z = 1$. A further step must be taken to approximate μ_Z and σ_Z as functions of n. The normalizing transformation of X (the original W or W' statistic) then becomes

$$Z' = \{(g(X) - \mu)|\sigma - \mu_z\}/\sigma_Z = (g(X) - \mu - \sigma\mu_Z)/(\sigma\sigma_Z)$$

that is, $g(X)$ has mean $\mu + \sigma\mu_Z$ and SD $\sigma\sigma_Z$. The W test for 3PL data, $\ln(\mathbf{y} - \gamma)$, may be treated this way; see Table 2 for details. Here, γ is estimated as $\tilde{\gamma}$ such that $\ln(\mathbf{y} - \tilde{\gamma})$ has zero sample skewness (Royston, 1992c). The parameter τ in Table 2 is estimated from the data as the sample SD of $\ln(\mathbf{y} - \tilde{\gamma})$.

Table 1. Normalizing transformations for the Shapiro–Wilk W and Shapiro–Francia W' statistics

Test	Range of n	Transformation (g)	u	Parameter	Coefficients	
W	4–11	$-\ln\{\gamma-\ln(1-W)\}$	n	γ	0	−2·273
					1	0·459
				μ	0	0·5440
					1	−0·39978
					2	0·025054
					3	−0·0006714
				$\ln(\sigma)$	0	1·3822
					1	−0·77857
					2	0·062767
					3	−0·0020322
W	12–2000[a]	$\ln(1-W)$	$\ln(n)$	μ	0	−1·5861
					1	−0·31082
					2	−0·083751
					3	0·0038915
				$\ln(\sigma)$	0	−0·4803
					1	−0·082676
					2	0·0030302
W'	5–5000	$\ln(1-W')$	$\ln(v)-v$[b]	μ	0	−1·2725
					1	1·0521
			$\ln(v)+2/v$[b]	σ	0	1·0308
					1	−0·26758

[a] Claimed by Royston (1992a) to be valid up to $n=5000$.
[b] Where $v=\ln(n)$.

Table 2. Normalizing transformation for the Shapiro–Wilk W statistic for data having a 3PL distribution

Range of n	Parameter smoothing expressions for the mean and SD of Z[a]
5–11	$\mu_z=-3{\cdot}8267+2{\cdot}8242u-0{\cdot}63673u^2-0{\cdot}020815v$
	$\sigma_z=-4{\cdot}9914+8{\cdot}6724u-4{\cdot}27905u^2+0{\cdot}70350u^3-0{\cdot}013431v$
12–2000	$\mu_z=-3{\cdot}7796+2{\cdot}4038u-0{\cdot}66756u^2+0{\cdot}082863u^3$ $-0{\cdot}0037935u^4-0{\cdot}027027v-0{\cdot}0019887vu$
	$\sigma_z=\ 2{\cdot}1924-1{\cdot}0957u+0{\cdot}33737u^2-0{\cdot}043201u^3$ $+0{\cdot}0019974u^4-0{\cdot}0053312vu$

[a] Where $u=f(n)=\ln(n)$, $v=u(\tau-\tau^2)$ and τ^2 is the variance of $\ln(y-\gamma)$.

3.3 Censored data

Verrill & Johnson (1988) have published tables of the 90th, 95th and 99th centiles of the square root of W and W' for $20 \leqslant n \leqslant 5000$ and $\Delta=0$, 0·2, 0·4, 0·6 and 0·8, based on about 5000 repetitions for each sample size and each value of Δ. The values for $n=5000$ are derived from the asymptotic distributions. For computer application, it is useful to obtain a P value for any combination of n and Δ within the above limits. The published data were used to estimate μ_Z and σ_Z as follows.

Define Z_α as the 100αth centile of the distribution of Z, the transformed W or W' statistic described in Sections 2.3 and 3.1. Empirical values of Z_α for $\alpha=0{\cdot}9$, 0·95 and 0·99 and $n=20$, 30, 40, 60, 80, 150, 250, 500 and 5000 were calculated from Verrill & Johnson's

Table 3. Smoothing constants for calculating R_α (see text for details)

	A_α		B_α		C_α	
α	W	W'	W	W'	W	W'
0·90	0·1640	0·1843	0·533	1·560	0·556	0·371
0·95	0·1736	0·1894	0·315	0·270	0·622	0·624
0·99	0·256	0·248	−0·00635	0·0	—	—

tables 1, 2 and 3 using the transformations given in Table 1. For $\Delta > 0$ and for each of the above values of n, the Z_α were approximated in terms of Δ as follows:

$$Z_\alpha = \Phi^{-1}(\alpha) + DR_\alpha^{-\ln(\Delta)}$$

The values of R_α were smoothed as functions of $u = \ln(n)$ as follows:

$$\begin{aligned} R_\alpha &= A_\alpha + B_\alpha C_\alpha^u \qquad \text{for } \alpha = 0{\cdot}9,\ 0{\cdot}95 \\ &= A_\alpha + B_\alpha u \qquad \text{for } \alpha = 0{\cdot}99 \end{aligned}$$

Estimates of A_α, B_α and C_α are given in Table 3. The values of D were approximately independent of α and were smoothed as

$$\begin{aligned} D &= 1 + 0{\cdot}8378u \qquad \text{for } W \\ &= 0{\cdot}76676u + 0{\cdot}015814u^2 \qquad \text{for } W' \end{aligned}$$

To use the method for given Z, n and Δ, the three values of Z_α are first found as just described. Then μ_Z and σ_Z are estimated as the intercept and slope, respectively, from the linear regression of Z_α on $\Phi^{-1}(\alpha)$.

3.4 *Worked examples*

The following are the values (in mmol/l) of fasting blood glucose from 24 type 1 diabetic patients (Thuesen *et al.* 1985; reproduced by Altman, 1991, table 11.6): 4·2, 4·9, 5·2, 5·3, 6·7, 6·7, 7·2, 7·5, 8·1, 8·6, 8·8, 9·3, 9·5, 10·3, 10·8, 11·1, 12·2, 12·5, 13·3, 15·1, 15·3, 16·1, 19·0, 19·5. Suppose that the 11 glucose values above 10 mmol/l had been recorded only as '> 10' (type I censoring). We wish to test the null hypothesis that the data were sampled from a normal distribution. For illustration, we calculate the P values from the W and W' tests for the complete and censored samples.

3.4.1 Complete sample. We have $n = 24$, $W = 0{\cdot}94525$. From Table 1, $g(W) = \ln(1 - W) = -2{\cdot}905$ and $u = \ln(n) = 3{\cdot}1781$. The mean μ and SD σ of $\ln(1 - W)$ are obtained using the polynomials in u given in Table 1:

$$\begin{aligned} \mu &= -1{\cdot}5861 - 0{\cdot}31082u - 0{\cdot}083751u^2 + 0{\cdot}0038915u^3 \\ &= -3{\cdot}295 \\ \ln(\sigma) &= -0{\cdot}4803 - 0{\cdot}082676u + 0{\cdot}0030302u^2 \\ &= -0{\cdot}7124 \end{aligned}$$

so $\sigma = \exp(-0{\cdot}7124) = 0{\cdot}4904$. We obtain $Z = \{-2{\cdot}905 - (-3{\cdot}295)\}/0{\cdot}4904 = 0{\cdot}795$. Finally, the P value of the W test is $1 - \Phi(Z)$, the tail area of the standard normal distribution to the right of Z, and equals $1 - \Phi(0{\cdot}795)$ or 0·21. This is not significant at the

5% level, so the null hypothesis of normality cannot be rejected. More briefly, for the Shapiro–Francia test we use values from the lower half of Table 1. We have $n=24$, $W'=0{\cdot}95484$, $g(W')=\ln(1-W')=-3{\cdot}0975$. For μ we have $u=\ln\ln(n)-\ln(n)=-2{\cdot}0218$, $\mu=-3{\cdot}3997$, and for σ, $u=\ln\ln(n)+2/\ln(n)=1{\cdot}7856$, $\sigma=0{\cdot}5530$, from which $Z=0{\cdot}546$ and $P=0{\cdot}29$ (not significant at the 5% level).

3.4.2 Censored sample. We have $n=24$, $\Delta=11/24=0{\cdot}4583$, $W=0{\cdot}92619$. As before, using Table 1 we have $g(W)=\ln(1-W)$, $u=\ln(n)=3{\cdot}1781$, $\mu=-3{\cdot}295$, $\sigma=0{\cdot}4904$. Thus $g(W)=-2{\cdot}606$ and $Z=\{g(W)-\mu\}/\sigma=1{\cdot}404$. We must evaluate the mean μ_z and SD σ_Z of Z, which are not now 0 and 1. Using the methods of Section 3.3, and the values of A_α, B_α and C_α given in Table 3, we have

$$\begin{aligned} R_{0\cdot9} &= A_{0\cdot9}+B_{0\cdot9}C_{0\cdot9}^{u}=0{\cdot}1640+0{\cdot}533(0{\cdot}556)^{u} \\ &= 0{\cdot}2465 \end{aligned}$$

Similarly, $R_{0\cdot95}=0{\cdot}2433$ and $R_{0\cdot99}=0{\cdot}2358$. Then $D=1+0{\cdot}8378u=3{\cdot}663$, so $Z_{0\cdot9}=\Phi^{-1}(0{\cdot}9)+3{\cdot}663\times0{\cdot}2465^{-\ln(11/24)}=2{\cdot}510$. Similarly, $Z_{0\cdot95}=2{\cdot}861$ and $Z_{0\cdot99}=3{\cdot}513$. The values $Z_{0\cdot9}$, $Z_{0\cdot95}$ and $Z_{0\cdot99}$ are respectively the 90th, 95th and 99th centiles of the null distribution of Z, so they may be used as critical values for significance testing of Z if a precise P value is not required. Here $Z<Z_{0\cdot9}$, so the test is not significant at the 10% level. For a more precise P value, regression of the Z_α on $\Phi^{-1}(\alpha)$ $(\alpha=0{\cdot}9,\ 0{\cdot}95,\ 0{\cdot}99)$ gives similar calculations with the W' test, we obtain $W'=0{\cdot}95382$, $Z=0{\cdot}587$, $\mu_Z=0{\cdot}891$, $\sigma_Z=0{\cdot}993$, $Z'=-0{\cdot}306$, $P=0{\cdot}62$. Thus no test indicates rejection of the null hypothesis of normality.

4 Comments

In theory, the methods described above should permit goodness-of-fit tests on data ranging from complete, continuous, putatively normal samples to grouped, singly censored data with an assumed lognormal distribution. Although not all such combinations have been examined, the P values from the W or W' tests should provide at least some guidance as to the acceptability of the distributional model. It should be stressed, however, that a normal probability plot should *always* accompany a test of non-normality, as the latter is only a summary statistic whereas the former shows all the data, including peculiarities not detected by the test. Furthermore, it should be realised that the requirement of normality is far more critical in some cases than in others. It is likely to be less critical for checking distributional assumptions behind test statistics derived from data aggregates, such as the t or F tests in analysis of variance, than when estimating quantiles of a distribution, such as for clinical reference ranges.

The choice between the W and W' tests is to some extent a matter of taste and convenience. W is more powerful than W' against short-tailed distributions such as the uniform, whereas W' is more sensitive to long-tailed symmetric alternatives such as the scale-contaminated normal. W' is the simpler to compute, but that is unimportant if your favourite package already gives W. Both tests may be recommended.

The doubly-censored normal distribution is not catered for, but since most of the information about non-normality lies in the tails of the sample, the power to detect departure from such a distribution in moderate-sized samples would be feeble indeed. Nevertheless, a normal plot for such data can be constructed in the usual way. Further research is needed to produce a test of departure from a mixture of normals. Since maximum likelihood fitting of the mixture model requires an iterative algorithm and the likelihood tends to be badly behaved, the simulations are likely to be computationally demanding.

The W test described here and in Royston (1992a) supersedes a previous version (Royston, 1982b, c; Royston 1983a). Investigation revealed that Shapiro & Wilk's (1965) approximation to the weights (**a**), used by Royston (1982b, c), was inadequate, therefore the earlier 'W test' differed seriously from the true test. I hope to publish an update to Royston's (1982c) algorithm incorporating the new method for W (Royston, 1993). The normalizing transformation for W' given here (Royston, 1992d) also supersedes an earlier one (Royston, 1983b), which was effective but cumbersome.

References

ALTMAN, D. G. (1991) *Practical Statistics for Medical Research*, p. 300 (London, Chapman and Hall).

DAVIES, C. S. & STEPHENS, M. A. (1978) Algorithm AS128. Approximating the covariance matrix of normal order statistics, *Applied Statistics*, 27, pp. 206–212.

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

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

ROYSTON, J. P. (1982c) Algorithm AS181. The W test for normality, *Applied Statistics*, 31, pp. 176–180.

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

ROYSTON, J. P. (1983b) A simple method for evaluating the Shapiro–Francia W' test for non-normality, *The Statistician*, 32, pp. 297–300.

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

ROYSTON, P. (1992b) Aspects of goodness-of-fit tests for normality. I. Use of the tests when the data contain ties, *Technometrics*, submitted.

ROYSTON, P. (1992c) Estimation, reference ranges and goodness-of-fit for the three-parameter lognormal distribution, *Statistics in Medicine*, 11, pp. 897–912.

ROYSTON, P. (1992d) A pocket-calculator algorithm for the Shapiro-Francia test of non-normality: an application to medicine, *Statistics in Medicine*, in press.

ROYSTON, P. (1993) The Shapiro–Wilk W test for non-normality in complete and singly censored samples, *Applied Statistics*, submitted.

SHAPIRO, S. S. & FRANCIA, R. S. (1972) An approximate analysis of variance test for normality, *Journal of the American Statistical Association*, 67, pp. 215–216.

SHAPIRO, S. S. & WILK, M. B. (1965) An analysis of variance test for normality (complete samples), *Biometrika*, 52, pp. 591–611.

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

THUESEN, L., CHRISTIANSEN, J. S. & FALSTIE-JENSEN, N. (1985) Increased myocardial contractility in short-term Type I diabetic patients: an echocardiographic study, *Diabetologia*, 28, pp. 822–826.

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