ON THE DISTRIBUTIONS OF THE W AND W' STATISTICS
FOR TESTING FOR NORMALITY

BY

M. A. STEPHENS

TECHNICAL REPORT NO. 218
APRIL 11, 1975

THIS RESEARCH WAS SPONSORED BY THE ARMY RESEARCH OFFICE
OFFICE OF NAVAL RESEARCH, AND AIR FORCE OFFICE OF
SCIENTIFIC RESEARCH BY CONTRACT NO.
N00014-67-A-0112-0085 (NR-042-267)

Reproduction in whole or in part is permitted
for any purpose of the United States Government
Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE DISTRIBUTIONS OF THE W AND W' STATISTICS FOR TESTING FOR NORMALITY

by

M. A. Stephens

Technical Report No. 218
April 11, 1975

Prepared under Contract N00014-67-A-0112-0085
(NR-042-267)
Office of Naval Research

Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted for any Purpose of the United States Government

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE DISTRIBUTIONS OF THE W AND W' STATISTICS
FOR TESTING FOR NORMALITY

by

M. A. Stephens

1. Introduction. The statistic W.

The W-statistic was introduced by Shapiro and Wilk (1965) as a
suitable statistic for testing for normality when the mean and variance
are not known. The statistic is based on the regression of the order
statistics of the sample on the expected values \( \sim m_i \), \( i = 1,2,\ldots,n \) of
the order statistics of a \( N(0,1) \) sample of size \( n \). Suppose \( m_i \)
are components of a vector \( \sim m \); let \( y_i \), \( i = 1,2,\ldots,n \), be the sample
values, and let these be components of a vector \( \sim y \). Shapiro and Wilk
(hereafter abbreviated SW) took the components of \( \sim m \) and of \( \sim y \) in
ascending order of size, but they may be placed both in descending order.
Let \( \sim v_{ij} \) be the covariance of the i- and j-order statistics from a
\( N(0,1) \) sample of size \( n \), and let \( \sim V \) be the covariance matrix with
elements \( v_{ij} \). Let

\[
\sim a = \sim m^\prime \sim V^{-1} m
\]

(1)

\[
\sim R^2 = \sim m^\prime \sim V^{-1} \sim m
\]

\[
\sim C^2 = \sim m^\prime \sim V^{-1} \sim V^{-1} \sim m
\]

where \( m^\prime \) denotes the transpose of \( m \).

The least squares estimate of \( \sigma \) is \( \hat{\sigma} = \sim m^\prime \sim V^{-1} \sim y/\sim m^\prime \sim V^{-1} \sim m \), and the
test statistic \( W \) is defined by
(2) \[ W = \frac{R^2 \delta^2}{C^2 S^2}; \]

i.e.

\[ \sqrt{W} = \frac{m'\gamma^{-1} \gamma}{(C \sqrt{\nu n-1})} \]

where \( S^2 = \sum \frac{(y_i - \bar{y})^2}{n} \) and \( s^2 = S^2/(n-1) \). The \( W \)-statistic was found to provide a good omnibus test for normality. For \( n > 20 \), however, \( \gamma \) is not exactly known, and SW gave approximations for the \( \gamma \) vector, and for \( R^2 \) and \( C^2 \). In addition, they show \( \max(W) = 1 \), and also

(3) \[ E(\sqrt{W}) = \frac{R^2 \Gamma(\frac{1}{2} n - \frac{1}{2})}{\sqrt{2} C \Gamma(\frac{1}{2} n)}; \quad E(W) = \frac{R^2 (R^2 + 1)}{C^2 (n-1)} \]

The exact null distribution of \( W \) is very difficult to find, and so, using the exact \( \gamma \) for \( n \leq 20 \) and the approximation for \( n > 20 \), SW derived Monte Carlo points for \( n \) up to 50. Thus the test is available for these values of \( n \). An extension was later made by Shapiro and Francia (1972; called SF); their test statistic is

(4) \[ W' = \frac{(m'\gamma)^2}{m'mS^2}; \]

so that, in \( W \), \( m'\gamma^{-1} \) has been replaced by \( m' \). Since values of \( m \) are available for very large sample sizes, \( W' \) can be used for \( n > 50 \), and SF gave Monte Carlo points for \( n \) up to 99.

In this paper we investigate the asymptotic distributions of \( W \) and of \( W' \). These would, if known, provide a useful supplement to the Monte Carlo values. The exact asymptotic distributions are not found,
but related results are given, and very good approximations are given for the significance points required for testing, based on a $\chi^2$ approximation to the distribution of $n(1 - \sqrt{W})$ or of $n(1 - \sqrt{W'})$.

2. Asymptotic results - moments.

We start by examining the limiting values of the expressions (3). The SW approximations for $R^2$ and $C^2$ suggest that $\lim_{n \to \infty} R^2/n = 2$ and that $\lim_{n \to \infty} C^2/n = 4$. This is heuristically demonstrated in Stephens (1973); it is also shown there that

$$m' \sqrt{W^{-1}} \sim 2m',$$

a result which gives support to the $W'$ statistic as a replacement for $W$, for large $n$. Following SW notation let $E(\sqrt{W}) = \mu^{1/2}$, and $E(W) = \mu$. Substitution of the above limits gives $\lim_{n \to \infty} \mu^{1/2} = 1$ and $\lim_{n \to \infty} \mu = 1$ (all limits will be as $n \to \infty$). Thus the mean of $W$ approaches its maximum. For finer approaches to the limits, we suppose that, for large $n$,

$$R^2 = 2 + \frac{a}{n} + \frac{b}{n^2}; \quad C^2 = 4 + \frac{c}{n} + \frac{d}{n^2}. \quad (6)$$

It is convenient to work with $m = n - 1$. Let $P = \Gamma(m/2)/\Gamma(n/2)$. $P$ can be expanded

$$P = \frac{\sqrt{2}}{m} \left( 1 + \frac{1}{4m} + \frac{1}{32m^2} + \ldots \right) \quad (7)$$

and after some algebra we have

3
\[ \mu^{1/2} = 1 + \frac{4a-c+6}{8m} + O\left(\frac{1}{m^2}\right) \]

and

\[ \mu = 1 + \frac{4a-c+6}{4m} + O\left(\frac{1}{m^2}\right). \]

Thus if \( \mu \) and \( \mu^{1/2} \) are plotted against \( 1/m \), one limiting slope is twice the other! This is clearly verified if one plots the estimates of \( \mu^{1/2} \) and \( \mu \), provided by SW from Monte Carlo studies, even though the estimates are approximate. This relation means that the variance of \( \sqrt{W} \), say \( \sigma^2 \), is of the form

\[ \sigma^2 = -\frac{2a+5}{8m} + O\left(\frac{1}{m^3}\right). \]

Thus to obtain an asymptotic distribution it appears that we need to consider \( z = m(\sqrt{W} - \mu^{1/2}) \) or \( n(\sqrt{W} - \mu^{1/2}) \).

3. Asymptotic results - distributions.

Consider first the statistic \( \tilde{W}^* = \tilde{m}'\tilde{Y}^{-1} - \frac{1}{\sqrt{n}}\tilde{m}\tilde{C}\sigma \); apart from a term whose limit is 1, this replaces \( s \) by \( \sigma \) in \( \sqrt{W} \). \( \tilde{W}^* \) is now a linear combination of order statistics, and Theorem 9.6 of David (1970) applies. Let \( T_n = \tilde{W}^* \). Using the result \( \tilde{m}'\tilde{Y}^{-1} - 2\tilde{m}' \) for large \( n \), and \( \lim \tilde{C}^2/n = 4 \), we have, following David's notation,

\[ \frac{1}{n} J(1/n) + \frac{m_1}{n}, \]
and so $J(i/n) \to m_i$. But for large $n$, $m_i$ is given by $F(m_i) = i/n$ where $F(\cdot)$ is the distribution function of the $y$-values; thus asymptotically $J(F(m_i)) = m_i$, and the limiting mean of $T_n$ becomes

$$
\mu_\infty = \int_{-\infty}^{\infty} x^2 \, dF(x)
$$

Here $F(x)$ is the normal distribution, so $\mu_\infty = 1$. David's $\sigma^2$ is then

$$
\sigma^2 = 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s \, F(s) \, t(1-F(t)) \, ds \, dt,
$$

which can be shown to be 0.5. The theorem then states that $\sqrt{n}(W^*-1)$ has the limiting normal distribution $N(0, 0.5)$.

Write $z$ for $s/\sigma$; then $W^* = \sqrt{W} \, z$ and

$$
\sqrt{n}(1-W^*) = \sqrt{n}(1-\sqrt{W})z + \sqrt{n}(1-z);
$$

i.e., if $u_1 = \sqrt{n}(1-W^*)$ and $u_2 = \sqrt{n}(1-W)$,

(10) \hspace{1cm} u_1 = u_2 \, z + \sqrt{n}(1-z).

Asymptotically, $u_1$ has mean 0 and variance 0.5, and so also does $\sqrt{n}(1-z)$; thus $u_2$ must converge in probability to zero. The asymptotic distribution of $u_2$ then poses an interesting characterization problem, as follows. SW show that $W$ and hence $u_2$ is independent of $z$, and $z$ has known distribution even for finite $n$. Thus the problem
is to find a density $f(u_2)$ for $u_2$, which when multiplied by the
density of $z$, transformation (10) made, and $z$ integrated out, gives
the large-$n$ distribution of $u_1$. However, to do this, knowledge of
the limiting distribution of $u_1$ is not enough; one needs to know also
the rate at which $u_1$ approaches $N(0,0.5)$.

It may be shown that, for the SF statistic $W'$,

$$E(\sqrt{W'}) = \frac{K\phi}{\sqrt{2}}; \quad E(W') = \frac{(D^2 + K^2)/(n-1)}{K^2}$$

where $K^2 = \bar{m}'\bar{m}$ and $D^2 = \bar{m}'\bar{m}$. If $W_1$ is defined as $\sqrt{W'}$ s.d.,
similar reasoning to that given for $W^*$ will give the asymptotic
distribution of $\sqrt{n}(1 - W_1) = N(0,0.5)$, and $n(1 - \sqrt{W'})$ has an asymptotic
distribution.

4. Related Statistics.

A statistic very similar to $W'$ is $r_n^2$ defined as follows. Let
$H(\cdot)$ be the inverse of the normal distribution function, let $H(i/(n+1))$
be called $H_i$, and let $H_i$, $i = 1,2,\ldots,n$, be the components of a
vector $H$. Define

$$r_n^2 = \frac{(H'h)^2}{H'HS}\hat{\tau}$$

essentially $\hat{\tau}$ has replaced $\tau$ in the definition of $W'$.
$r_n^2$ has been
investigated by de Wet and Venter (1972); we have closely followed their
notation in the definition above, but have replaced their $H_{in}$ by $H_i$.
de Wet and Venter show that the limiting distribution of the statistic
$2n(1-r_n^2 - a_n) = Z$, say, is the same as that of $\sum_{r=3}^{n} (Y_r^2 - 1)/r$, where
\( Y_r \) are independent \( \mathcal{N}(0,1) \) variables. Values of \( a_n \) depend on the \( \bar{H} \) vector, and on a vector whose components are the derivative \( \frac{dH(y)}{dy} \) evaluated at \( y = i/(n+1) \); also on a matrix \( T \) closely related to \( Y \). \( T \) has entries
\[
T_{ij} = \frac{i}{n+1} \left( 1 - \frac{j}{n+1} \right)
\]
for \( i \leq j \) and is symmetric. Values \( a_n \) have been tabulated by de Wet and Venter, and also values of the limiting distribution of \( Z \) given above.

Many results have been produced in this interesting paper concerning \( r_n^2 \) and related statistics, but it is not easy to translate these rigorously into results for \( W' \) or \( W \). For instance, the asymptotic results for \( r_n \) give support to the above reasoning that we should look for an asymptotic distribution for \( n(1 - \sqrt{W'}) \); then if it were true that \( n(\sqrt{W'} - r_n) \to b \) in probability as \( n \), we should have the result that the limiting distribution of \( 2n(1 - \sqrt{W'}) - a_n - b \) is also that of \( Z \). This seems difficult to prove, though de Wet and Venter get good results, when they make some comparisons between approximate points for \( r_n \), calculated from their asymptotic results, and Monte Carlo values for \( W' \) given by Shapiro and Francià. In these comparisons it appears that \( b \) is taken as zero.

5. Asymptotic results: numerical.

We shall leave the theoretical problems here and turn to numerical approximations. The evidence above supports that \( n(1 - \sqrt{W}) \) and \( n(1 - \sqrt{W'}) \) might be well approximated by \( \chi^2 \) distributions of the form \( rX^2_p \). We determine \( r \) and \( p \) by equating moments. The mean of \( n(1 - \sqrt{W}) \) will
be, using (8), \(- (4a - c + 6)/8\), and the variance will be \(- (2a + 5)/8\),
to be set equal to \(rp\) and \(2r^2p\) respectively.

Stephens (1973) has found estimates of \(a, b, c, d\); \(\hat{a} = -2.835\),
and \(\hat{c} = -0.812\), so that the estimated mean becomes 0.56, and estimated
variance 0.08. These lead to values \(r = 0.07\) and \(p \approx 8\). For prac-
tical utility, \(p\) should be an integer, so the approximation suggested
becomes \(n(1 - \sqrt{W}) = 0.07 \chi^2_8\). Asymptotically, \(W\) and \(W'\) approach
each other, and one might speculate that the same \(\chi^2\) approximation will
be valid for \(W'\). However, these approximations depend on the way in
which asymptotic limits are approached, so we shall treat \(W'\) as a separate
problem. We use the well-known result \(\lim_{n \to \infty} \frac{m'm}{n} = 1\); also,
\(\lim_{n \to \infty} \frac{m'y_m}{n} = 0.5\) from results in Stephens (1973). Then, using coefficients
from expansions

\[
\left(12\right) \quad \frac{D^2}{n} = \frac{1}{2} + \frac{a}{n} + \frac{b}{n^2} \, ; \quad \frac{K^2}{n} = 1 + \frac{c}{n} + \frac{d}{n^2}
\]

we can find limiting forms for the moments:

\[
E(\sqrt{W'}) = 1 + \frac{2c+3}{4n} + O\left(\frac{1}{n^2}\right),
\]

\[
E(W') = 1 + \frac{2c+3}{2n} + O\left(\frac{1}{n^2}\right),
\]

and finally variance \((\sqrt{W'}) = (a - c - 5/8)/n^2\). Results from least
squares fitting give \(\hat{a} = -1.795\), \(\hat{c} = -2.530\) (with \(\hat{b} = 4.076\);
\(\hat{d} = 4.458\)), and so \(Var(\sqrt{W'})\) can be estimated as \(0.11/n^2\). The mean
and variance of \(n(1-\sqrt{W})\) are then approximately 0.51 and 0.11.
Using an \(rX^2_p\) approximation, we have \(r = 0.108\) and \(p \approx 5\). The
suggested approximation is therefore \(n(1-\sqrt{W'}) = 0.108 X^2_5\).

(a) Statistic $W'$. We first compare the percentage points of $W'$, given by the approximation above, with those given by a Monte Carlo study reported in SF. For $n = 50, 75, 100$ the values are given in Table 1. It is clear that the approximation is very good indeed for the lower tail of $W' -- the tail which is used for the test. The upper tail is not good, as often happens when this simple type of approximation is used for a very steep tail.

(b) Statistic $W$. It is more difficult to verify the accuracy of the approximation $n(1-\sqrt{W}) = 0.07 \chi^2_8$, given in §4. This is because SW gave Monte Carlo points for $W$, when the statistic is calculated using an approximation for $\tilde{a}'$ and not the correct $\tilde{a}'$. Although this approximation is a very good one, it has an effect; this can be seen by plotting the Monte Carlo estimates of $\mu^{1/2}$ and $\mu$ against $1/n$. For large $n$, they lie on a curve which is almost a straight line, and when this is extrapolated to give values for $n \to \infty$, the values are $\mu^{1/2} = .992$ and $\mu = .985$ instead of 1 for each. Evidently the effect of the approximate coefficients is to give values on the whole too small. Thus an approximation on the lines of $n(1-K\sqrt{W}) = r\chi^2_p$, with $K$ slightly more than 1, might be better for the percentage points of the $W$ as calculated. In fact, $n(1-1.003\sqrt{W}) = .076 \chi^2_8$ gives good results for $n = 50$: the approximation gives $.954, .948, .934$ for the lower tail 10, 5, and 1 percentage points, is compare with SW Monte Carlo values: $.954, .948, .935$. For $n = 45$ the corresponding figures are $.950, .943, .928$ and SW Monte Carlo values are $.952, .943, .929$. This is as good as a more complicated
approximation given by Shapiro and Wilk (1968) from which the Monte Carlo results have been quoted. Nevertheless, the \( \chi^2 \) approximation drops in accuracy for smaller \( n \), and is not suggested as a working approximation but more as a guide to the accuracy of the ideas leading to these approximations.

7. Further remarks.

The inspiration of this work, as noted above, was an attempt to add to the theory of \( W \) and \( W' \); in particular, they have interesting asymptotic distributions which it would be useful to know. For this to proceed, we need firstly, information on the rate at which \( u_1 = \sqrt{n} (1-W^*) \), and the corresponding \( \sqrt{n} (1-W_1) \) approach their asymptotic normal distributions; and secondly, knowledge of the large-\( n \) expansions of \( R^2/n \), \( C^2/n \), \( D^2/n \) and \( K^2/n \). Of these, only \( K^2 = m'm \) has been much investigated (see, for example, Ruben (1956) and Saw and Chow (1966) for some interesting formulas for \( m'm \)), but both problems have considerable intrinsic interest for work concerning normal order statistics. On the practical side, approximations for the finite-\( n \) points for \( W \) would not be of much value, in terms of making the statistic easy to use, unless at the same time good methods could be found to calculate the \( a_i \) coefficients given in SW's tables. However, for \( W' \), we have given a good simple approximation for the percentage points used in testing, for \( n \) at least as low as 50, and values of \( m_i \), needed for the calculation of \( W' \), can be found from approximations given by Harter (1961); thus the whole \( W' \) test can be put on a computer without needing to read in coefficients. For large samples, \( W' \) appears to provide a good test for a wide range of alternatives. (SF: see also Stephens (1974).)
TABLE 1

Percentage points for the SF statistic: comparison of Monte Carlo values and those given by the $\chi^2$ approximation in Section 5

<table>
<thead>
<tr>
<th>Sample size n:</th>
<th>Percentage level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>MC</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>75</td>
<td>MC</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>100</td>
<td>MC</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>
References


ON THE DISTRIBUTIONS OF THE W AND W' STATISTICS FOR TESTING FOR NORMALITY

M. A. STEPHENS

Department of Statistics
Stanford University
Stanford, Calif. 94305

Office of Naval Research
Statistics & Probability Program Code 436
Arlington, Virginia 22217

April 11, 1975

OFFICIAL USE ONLY

Approved for public release: Distribution unlimited.

Shapiro-Wilk Statistic
Shapiro-Francia Statistic Tests for normality
normal order statistics

An investigation is made into the asymptotic distributions of W and W', statistics introduced by Shapiro and Wilk and Shapiro and France for testing for normality. Some simple approximations are given for the null percentage points for large n. Some problems are posed concerning asymptotic behavior of several scalar quantities of interest in the study of normal order statistics.