18.465, September 12, 2012
R. Dudley

THE SHAPIRO–WILK TEST FOR NORMALITY

Given a sample $X_1, \ldots, X_n$ of n real-valued observations, the Shapiro–Wilk test (Shapiro and Wilk, 1965) is a test of the composite hypothesis that the data are i.i.d. (independent and identically distributed) and normal, i.e. $N(\mu, \sigma^2)$ for some unknown real μ and some $\sigma > 0$.

This test of a parametric hypothesis relates to nonparametrics in that a lot of statistical methods (such as t-tests and analysis of variance) assume that variables are normally distributed. If they are not, then some nonparametric methods may be needed.

The 2-parameter normality hypothesis cannot simply be reduced to a simple hypothesis. Of course, the variables $X_i - \mu$ are i.i.d. $N(0, \sigma^2)$ and $(X_i - \mu)/\sigma$ are i.i.d. $N(0, 1)$, but these variables are not observed because μ and σ are unknown. If we replace μ by its usual estimate $\overline{X} = (X_1 + \cdots + X_n)/n$ and consider $X_i - \overline{X}$, then these variables have the same distribution, which is normal with mean 0, but they are dependent (their sum is 0). If we replace σ by the usual estimate

$$s_X = \left(\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2 \right)^{1/2},$$

then $\sqrt{n}(\overline{X} - \mu)/s_X$ has a t_{n-1} distribution (but involves μ), and $(X_i - \overline{X})/s_X$ don't have even that nice a distribution and are still more dependent.

1. Classical diagnostics for non-normality: skewness and kurtosis

Long before the Shapiro–Wilk test (or any other such general test) for normality was invented, statisticians used the following diagnostics. For a random variable X with $E(|X|^3) < \infty$, mean $EX = \mu$, and standard deviation $\sigma > 0$, the *skewness* of X or its distribution is defined as $\gamma_1(X) = E((X - \mu)^3/\sigma^3)$. Since any normal distribution is symmetric around its mean μ, its skewness is 0. For example, if Z has standard normal distribution $N(0, 1)$ then $EZ^3 = 0$. The skewness is unchanged if we add any constant to X or multiply it by any positive constant. The skewness can have any real value.

An integration by parts shows that $E(Z^4) = 3$. For any random variable X with $E(X^4) < \infty$, mean $EX = \mu$, and standard deviation $\sigma > 0$, the *kurtosis* is defined by

$$\gamma_2(X) \;=\; \frac{E((X - \mu)^4)}{\sigma^4} - 3.$$

It's also sometimes called "excess kurtosis" or just "excess." The kurtosis is unchanged if we add a constant to X or multiply it by any non-zero constant. Any normal distribution or random variable has 0 kurtosis.

The kurtosis is clearly larger than -3. (In fact, it's always at least -2; to see this we can assume $\mu = 0$, and then $E(X^4) \geq \sigma^4$ because $\operatorname{Var}(X^2) \geq 0$. Letting $X = \pm 1$ with probability $1/2$ each we see that $\gamma_2(X) = -2$, the smallest possible value.)

Forms of skewness and kurtosis for finite samples $(X_1, \ldots, X_n)$ are defined by replacing μ by $\overline{X}$, σ by the sample standard deviation s_X, X by X_j, and E by $\frac{1}{n}\sum_{j=1}^{n}$. The sample

skewness and kurtosis are defined for any finite sample with $s_X > 0$. If one of them is notably different from 0, and n is fairly large, it appears that the observations may not be normally distributed.

2. The Shapiro–Wilk test: Basics of Use in R

In practice, the test is simple to apply on a computer using R. Namely, let $X = (X_1, \ldots, X_n)$ be the data vector, represented in R if entered individually as $c(X_1, \ldots, X_n)$. Type

shapiro.test(X)

and you will see as output a test statistic called W (for Wilk) and a p-value. If the p-value is less than, say, the conventional level 0.05, then one rejects the normality hypothesis, otherwise one doesn't reject it. To apply the test it isn't necessary at first to understand W, but in this course we're going to try. It always satisfies $0 < W \leq 1$. For values of W close enough to 1 (depending on n) the normality hypothesis will not be rejected. For smaller W it will be rejected.

For $n = 2$, normality can never be rejected, so the test is useful only for $n \geq 3$. The R implementation allows n up to 5,000.

3. Order Statistics

The sample skewness and kurtosis are each one-dimensional quantities. Together with $\overline{X}$ and s_X^2 they give a useful four-dimensional summary about the location, scale, and shape of the data distribution, but they don't completely characterize the data, as different data sets can have the same values of the four quantities. Whereas, if we arrange the data $X_1, \ldots, X_n$ in order, to get the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, we haven't lost any information. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0,1)$ and also take their order statistics $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$. Let's consider the expectations of $Z_{(j)}$, which of course depend on n, $m_j := m_{n,j} := EZ_{(j)}$ for the given n. A next idea is to consider the correlation of $X_{(1)}, X_{(2)} \ldots, X_{(n)}$ with $(m_1, m_2 \ldots, m_n) = (EZ_{(1)}, EZ_{(2)}, \ldots, EZ_{(n)})$, in other words, to ask whether the order statistics of X_j are well correlated with expected standard normal order statistics. A correlation close to 1 would suggest a good fit to normality, whereas a correlation much less than 1 would suggest non-normality. This idea is on the right track in that if we add a constant to all the X_j we will add the same constant to their order statistics and to $\overline{X}$, leaving $X_{(j)} - \overline{X}$ and s_X unchanged. Likewise if we multiply all X_j by a positive constant, the ratios $(X_{(j)} - \overline{X})/s_X$ will be unchanged and so will the correlation. (Both the ratios and the correlations are dimensionless.) Thus if the X_j are indeed i.i.d. normal, the correlation will have the same distribution, not depending on the location μ or scale σ of the X_j.

The $Z_{(j)}$ and their expectations m_j have some symmetry properties. The random variables $-Z_1, \ldots, -Z_n$ are also i.i.d. $N(0,1)$, but the ordering of these variables is reversed. Thus $Z_{(1)}$ has the same distribution as $-Z_{(n)}$, and more generally

$$\{-Z_{(n+1-k)}\}_{k=1}^n =_d \{Z_{(k)}\}_{k=1}^n \tag{1}$$

where "$=_d$" means that the two vector random variables are equal in distribution. It follows in particular that

$$m_j \;=\; -m_{n+1-j}, \quad j = 1, \ldots, n. \tag{2}$$

The squared correlation of $X_{(j)}$ with m_j, which I'll call W', gives a test statistic for normality called the Shapiro–Francia statistic (Shapiro and Francia, 1972). But the better known Shapiro–Wilk statistic uses not only the means but the covariances of the normal order statistics $Z_{(j)}$. (In fact, $Z_{(j)}$ is rather strongly correlated with its neighbors $Z_{(j-1)}$ and $Z_{(j+1)}$ because of the ordering $Z_{(j-1)} \le Z_{(j)} \le Z_{(j+1)}$.)

Let V be the $n \times n$ covariance matrix of the $Z_{(j)}$, $V_{ij} = E[(Z_{(i)} - m_i)(Z_{(j)} - m_j)]$. This is a positive definite symmetric matrix. For a given n let m be the $n \times 1$ column vector $(m_1, \ldots, m_n)'$, consider the vector $m'V^{-1}$ whose length is $C := (m'V^{-1}V^{-1}m)^{1/2}$, and let $a' := (a_1, ..., a_n) := m'V^{-1}/C$, which is a unit row vector. The Shapiro–Wilk statistic is then defined by

$$W \;=\; \left(\sum_{j=1}^{n} a_j X_{(j)}\right)^2 \Big/ \left(\sum_{j=1}^{n} (X_j - \overline{X})^2\right), \tag{3}$$

as in the paper of Shapiro and Wilk (1965).

They point out that the statistic is preserved by a change in location, adding a constant b to all the X_j and thus to each $X_{(j)}$. Since b will also be added to $\overline{X}$, this clearly won't change the denominator. For the numerator, it follows from (2) that $\sum_{j=1}^{n} m_j = 0$. From (1) it follows that the covariance matrix V is also preserved if we interchange j with $n+1-j$ in the indices of both rows and columns, and so the same will be true for V^{-1}. From this and (2), we have

$$a_j = -a_{n+1-j}, \quad j = 1, \ldots, n, \tag{4}$$

from which it follows that $\sum_{j=1}^{n} a_j = 0$, and that the numerator and so W are also unchanged by adding b to all X_j. If all X_j and so all $X_{(j)}$ are multiplied by a constant $c > 0$, then clearly the numerator and denominator of W are both multiplied by c^2, so W is again unchanged and is preserved by changes of scale.

If $n = 2k + 1$ is odd then (4) implies that the coefficient a_{k+1} of the sample median $X_{(k+1)}$ in the numerator is 0. Whether n is even or odd, the numerator of W can be written as

$$\left(\sum_{j=1}^{\lfloor n/2 \rfloor} -a_j (X_{(n+1-j)} - X_{(j)})\right)^2 \tag{5}$$

where $\lfloor x \rfloor$ is the largest integer $\le x$. (The minus sign in front of a_j makes no difference because of the squaring, but is given because in fact a_j for $2j \le n$ are negative.)

Here are some remarks on the weighting of different order statistics in the numerator of W. Suppose we were sampling from a heavy-tailed distribution such as a Cauchy distribution with density $f(x) = 1/(\pi(x-m)^2 + 1)$ for all real x, where m is a location parameter around which f is symmetric and so which is the median of the distribution. Such a distribution tends to produce outliers, namely, observations that are much larger than or much less than the other observations. Thus in estimating m, we would want

to downweight the extreme order statistics $X_{(1)}$ and $X_{(n)}$ which may be outliers. In particular, the sample mean $\overline{X}$ has the same distribution as an individual observation X_1 and so is a very ineffective estimator of m. The sample median, namely $X_{(k+1)}$ for $n = 2k+1$ odd, or $(X_{(k)} + X_{(k+1)})/2$ for $n = 2k$ even, is a much more effective estimator, although it ignores all the order statistics except the one or two in the middle.

For normal variables the situation is quite different. The numerator of W, omitting the squaring, was chosen by taking an efficient unbiased estimator of σ as a linear function of order statistics. For $n = 7$, for example, Shapiro and Wilk (1965, Table 5) give the coefficients $a_7 = -a_1 \doteq 0.6233$, $a_6 = -a_2 \doteq 0.3031$, $a_5 = -a_3 \doteq 0.1401$, and $a_4 = 0$. Thus the weight is highest for the most extreme order statistics and decreases as one goes inward toward the median. Such a pattern holds generally. It clearly also holds for m_j in place of a_j.

4. Consistency

A test of a given hypothesis, in this case normality, is said to be *consistent* against a given alternative P, namely that the X_j are i.i.d. but with some non-normal distribution P, if for any $\alpha > 0$, if we reject normality at level α, then the probability that we reject it approaches 1 as $n \to \infty$. A test is called simply *consistent* if it is consistent against all alternatives, in other words in this case, for any non-normal P and fixed $\alpha > 0$, normality will be rejected with probability converging to 1 as n becomes large.

Consistency of the Shapiro–Wilk test against all non-normal alternatives was conjectured in the 1970's. Sarkadi (1975) gave a proof for the Shapiro–Francia test. The proof is actually written for alternatives having a finite second moment. Sarkadi wrote "the author will prove in a subsequent paper" [presumably Sarkadi (1981)] "that this restriction is not essential." The consistency is more difficult for the Shapiro–Wilk test because the covariance matrices V of the normal order statistics and the inverses V^{-1} are not explicitly known. Leslie, Stephens, and Fotopoulos (1986) give a proof of consistency, while giving credit to Theorem 1 of Sarkadi (1981).

5. Values and Null Distribution of W

It was mentioned before that always $0 < W \leq 1$. In fact, for a given n, the possible values of W are bounded below by a strictly positive number $na_1^2/(n-1)$ according to Shapiro and Wilk (1965, Lemma 3). If the null hypothesis of normality holds, Shapiro and Wilk gave an exact distribution of W only for $n = 3$. In that case $a_1^2 = 1/2$ so the lower bound of possible values is $W \geq 3/4$. Shapiro and Wilk (1965, Corollary 4) give the distribution as a truncated Beta$(1/2, 1/2)$ distribution, namely, having density $(3/\pi)(1-w)^{-1/2}w^{-1/2}$ for $3/4 \leq w < 1$ and 0 elsewhere.

For $n \leq 20$, values of $V^{-1}m_i$ were available from Sarhan and Greenberg (1956, Table II, for $r_1 = r_2 = 0$ [complete samples], second line for each n [coefficients for estimating σ]), given to 8 decimal places. These could then easily be normalized to get the coefficients a_i, as given to 4 places in Shapiro and Wilk (1965, Table 5). For $20 < n \leq 50$, values of a_i are also given, but these were approximate. Then Shapiro and Wilk (1965, Table 6) gave percentage points for selected levels $\alpha = 0.01,\ 0.02, 0.05,\ 0.1,\ 0.5$, and $1-\alpha$ for those α, based on approximations and Monte Carlo simulations.

Instead of distributing copies of such tables, I suggest that we just use the R software to find p-values (for problem sets; for exams, questions will be asked only about theoretical properties, or with more specific information provided if needed).

Leslie, Stephens, and Fotopoulos (1986) found the limiting distribution of W as $n \to \infty$ when the null hypothesis of normality holds. Namely, they show that the distribution of $n(W - EW)$ converges to that of $-\zeta = -\sum_{k=3}^{\infty}(Z_k^2 - 1)/k$ where Z_k are i.i.d. $N(0,1)$ random variables. More specifically, they define a sequence of constants which they call a_n in equation (3) of their paper, but which are not the same as the coefficients a_n used in forming W, defined just before (3), so I'll call A_n what they call a_n. Leslie et al. state that they themselves, and Verrill and Johnson, in earlier unpublished technical reports, had shown for the Shapiro–Francia statistic W' that

$$\zeta_n := 2n(1 - \sqrt{W'}) - A_n \tag{6}$$

converges in distribution to ζ. Leslie et al. (1986) show in their Lemma (p. 1499) part (iv) that for some constants $C_j > 0$, $j = 1, 2$,

$$C_1 \log\log n < A_n < C_2 \log\log n, \tag{7}$$

and in part (ii) that $A_n - nE(1 - W) \to 0$ as $n \to \infty$. But the paper only shows convergence at a painfully slow rate. For example, inequality (1) of the paper gives a $(\log n)^{-1/2}$ rate, and the Lemma, part (iii), a $(\log n)^{-1}$ rate. For such reasons, the limit distribution seems not to be of much practical use, and for $n > 50$, Monte Carlo simulation seems still to be needed. For the maximum $n = 5{,}000$ that R allows, the simulation would have needed rather heavy computation. For that n, $(\log n)^{-1/2} \doteq 0.343$, not very small.

The representations given by Leslie et al. *are* useful in comparing different test statistics for normality. Namely, they show that under the normality hypothesis, the Shapiro–Wilk and Shapiro–Francia statistics W and W' are asymptotically equivalent in the sense (their equation (5)) that as $n \to \infty$, $n(\sqrt{W} - \sqrt{W'}) \to 0$ in probability. In other words, using the o_p notation (as in the "Convergence and boundedness in probability" handout), we have $\sqrt{W} - \sqrt{W'} = o_p(1/n)$. Since $0 < W \leq 1$ and $0 \leq W' \leq 1$ we always have $0 < \sqrt{W} + \sqrt{W'} \leq 2$. Multiplying this sum by the difference gives $W - W' = o_p(1/n)$ also.

Rewriting (6) gives

$$\sqrt{W'} = 1 - \frac{A_n}{2n} - \frac{\zeta_n}{2n} \tag{8}$$

where ζ_n converges to ζ in distribution. Squaring both sides of (8) gives

$$W' = 1 - \frac{A_n}{n} - \frac{\zeta_n}{n} + o_p\left(\frac{1}{n}\right) \tag{9}$$

because terms with n^2 in the denominator and at most quantities of order $(\log\log n)^2$ in the numerator are $o_p(1/n)$. Since $W - W'$ is also $o_p(1/n)$, (9) holds with the Shapiro–Wilk statistic W in place of W'. Moreover, since $\zeta_n + o_p(1)$ converges in distribution to ζ also, we in a sense don't need the $o_p(1/n)$ term as long as we bear in mind that ζ_n may be different by terms of order $o_p(1)$ in representations of the different statistics.

Further, Verrill and Johnson (1987) showed that three other test statistics for normality proposed by different authors, also of the form of squared correlations of some coefficients b_j with $X_{(j)}$, are all asymptotically equivalent in the same sense under the normality assumption. The Shapiro–Wilk statistic is perhaps the hardest to compute if one needs to (for example, if $n > 5{,}000$) although easily available in R for $n \leq 5{,}000$. The Shapiro–Francia and at least some of the others are easier, and so might be preferred for $n > 5{,}000$. Apparently only the Shapiro–Wilk test is implemented in R. I found that the Shapiro–Francia test has been implemented, not in MATLAB itself, but by someone on an "Open Exchange" site that MATLAB maintains.

For finite n, the relative power of the tests against given alternatives depends on n and the alternative. Filliben (1975) gave power comparisons against 52 alternatives for $n = 20$ and 50. Just comparing the W (Shapiro–Wilk) and W' (Shapiro–Francia) tests, Filliben found that the W test is more powerful than the W' against "symmetric alternatives shorter-tailed than normal" such as the uniform, whereas W' has slightly higher power against symmetric alternatives longer-tailed than normal such as the t_2 distribution, but slightly lower power against skewed alternatives.

6. Summary of asymptotic properties of the Shapiro–Wilk and related statistics

How does the Shapiro–Wilk statistic $W = W_n$ based on an i.i.d. sample behave as $n \to \infty$, first, if the samples are from a normal distribution? We can read off properties from (9), as follows:
(i) W_n converges to 1 in probability;

How fast?

(ii) $1 - W_n = O_p(A_n/n)$ where A_n grow at the rate $\log \log n$ by (7);

(iii) In more detail, we have (9) itself for W in place of W'.

Now, what if the X_j are i.i.d. with a non-normal distribution? In that case, not even (i) will hold, i.e. W_n will not converge to 1 in probability. If X_j have finite variance, then W_n will converge in probability to some number ρ^2 less than 1. If the distribution of X_j is close to normal, ρ^2 might be not far from 1, such as 0.985, but W_n will "get stuck" around that value and not get closer to 1 when n gets large. So in a sense, consistency is easier than showing that the different statistics have the same detailed asymptotic behavior in the normal case.

NOTES ON THE LITERATURE

If one knows a basic paper on a subject, such as that of Shapiro and Wilk (1965) in this case, one can look for more about it by doing a citation search, as with Web of Science. I found that as of early September 2010 there had been over 3400 published papers citing Shapiro and Wilk's (Google Scholar said 3116, and 5270 through Sept. 10, 2012, although it often gives more citations than Web of Science does). The great majority of the citers are just applying the test and so seemed not so interesting theoretically. In 2010 I scanned 200 relatively recent citations (during 2009). The most interesting to me

were a few papers on testing for multivariate normality, which we may or may not get to later in the course.

About the paper by Romão et al. (2010), in a journal that MIT libraries don't subscribe to, online I was only able to access an abstract and first page, which say that 33 different tests of normality are compared but otherwise is not very specific.

REFERENCES

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics* **17**, 111-117.

Leslie, J. R., Stephens, M. A., and Fotopoulos, S. (1986). Asymptotic distribution of the Shapiro–Wilk W for testing for normality. *Ann. Statist.* **14**, 1497–1506.

*Romão, X., Delgado, R.,and Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *J. Statistical Computation and Simulation* **80**, 545-591.

Sarhan, A. E., and Greenberg, B. G. (1956), Estimation of location and scale parameters by order statistics from singly and doubly censored samples, *Ann. Math. Statist.* **27**, 427-451.

Sarkadi, K. (1975), The consistency of the Shapiro–Francia test, *Biometrika* **62**, 445–450.

*Sarkadi, K. (1981), On the consistency of some goodness of fit tests, *Proc. Sixth Conf. Probab. Theory, Brasov, 1979*, Ed. Acad. R. S. Romania, Bucuresti, 195–204.

Shapiro, S. S., and Francia, R. S. (1972), An approximate analysis of variance test for normality, *J. Amer. Statist. Assoc.* **67**, 215-216.

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

Verrill, Steve, and Johnson, R. A. (1987), The asymptotic equivalence of some modified Shapiro–Wilk statistics — complete and censored sample cases, *Ann. Statist.* **15**, 413-419.

* – At this writing, I have not seen these papers, but in one case a reference to the paper in another source, and in another case an abstract only.