swilk — Shapiro–Wilk and Shapiro–Francia tests for normality

Description

`swilk` performs the Shapiro–Wilk W test for normality for each variable in the specified varlist. Likewise, `sfrancia` performs the Shapiro–Francia W' test for normality. See [MV] **mvtest normality** for multivariate tests of normality.

Quick start

Shapiro–Wilk test of normality

Shapiro–Wilk test for `v1`

```
swilk v1
```

Separate tests of normality for `v1` and `v2`

```
swilk v1 v2
```

Generate new variable `w` containing W test coefficients

```
swilk v1, generate(w)
```

Specify that average ranks should not be used for tied values

```
swilk v1 v2, noties
```

Test that `v3` is distributed lognormally

```
generate lnv3 = ln(v3)
swilk lnv3
```

Shapiro–Francia test of normality

Shapiro–Francia test for `v1`

```
sfrancia v1
```

Separate tests of normality for `v1` and `v2`

```
sfrancia v1 v2
```

As above, but use the Box–Cox transformation

```
sfrancia v1 v2, boxcox
```

Specify that average ranks should not be used for tied values

```
sfrancia v1 v2, noties
```

Menu

swilk

Statistics > Summaries, tables, and tests > Distributional plots and tests > Shapiro-Wilk normality test

sfrancia

Statistics > Summaries, tables, and tests > Distributional plots and tests > Shapiro-Francia normality test

Syntax

Shapiro–Wilk normality test

`swilk` *varlist* [*if*] [*in*] [, *swilk_options*]

Shapiro–Francia normality test

`sfrancia` *varlist* [*if*] [*in*] [, *sfrancia_options*]

swilk_options	Description
Main	
`generate(`*newvar*`)`	create *newvar* containing W test coefficients
`lnnormal`	test for three-parameter lognormality
`noties`	do not use average ranks for tied values

sfrancia_options	Description
Main	
`boxcox`	use the Box–Cox transformation for W'; the default is to use the log transformation
`noties`	do not use average ranks for tied values

`by` is allowed with `swilk` and `sfrancia`; see [D] **by**.

Options for swilk

Main

`generate(`*newvar*`)` creates new variable *newvar* containing the W test coefficients.

`lnnormal` specifies that the test be for three-parameter lognormality, meaning that $\ln(X-k)$ is tested for normality, where k is calculated from the data as the value that makes the skewness coefficient zero. When simply testing $\ln(X)$ for normality, do not specify this option. See [R] **lnskew0** for estimation of k.

`noties` suppresses use of averaged ranks for tied values when calculating the W test coefficients.

Options for sfrancia

Main

`boxcox` specifies that the Box–Cox transformation of Royston (1983) for calculating W' test coefficients be used instead of the default log transformation (Royston 1993a). Under the Box–Cox transformation, the normal approximation to the sampling distribution of W', used by `sfrancia`, is valid for $5 \le n \le 1000$. Under the log transformation, it is valid for $10 \le n \le 5000$.

`noties` suppresses use of averaged ranks for tied values when calculating the W' test coefficients.

Remarks and examples

stata.com

`swilk` can be used with $4 \le n \le 2000$ observations. `sfrancia` can be used with $10 \le n \le 5000$ observations; however, if the `boxcox` option is specified, it can be used with $5 \le n \le 1000$ observations.

Also see [R] **sktest** for the skewness and kurtosis test described by D'Agostino, Belanger, and D'Agostino (1990) with the empirical correction developed by Royston (1991b). While the Shapiro–Wilk and Shapiro–Francia tests for normality are, in general, preferred for nonaggregated data (Gould and Rogers 1991; Gould 1992b; Royston 1991b), the skewness and kurtosis test will permit more observations. Moreover, a normal quantile plot should be used with any test for normality; see [R] **diagnostic plots** for more information.

▷ Example 1

Using our automobile dataset, we will test whether the variables `mpg` and `trunk` are normally distributed:

```
. use http://www.stata-press.com/data/r15/auto
(1978 Automobile Data)

. swilk mpg trunk

                   Shapiro-Wilk W test for normal data

    Variable        Obs       W           V         z       Prob>z
                                                                     
         mpg         74    0.94821      3.335     2.627    0.00430
       trunk         74    0.97921      1.339     0.637    0.26215

. sfrancia mpg trunk

                  Shapiro-Francia W' test for normal data

    Variable       Obs       W'          V'        z       Prob>z
                                                                    
         mpg        74    0.94872      3.650     2.510    0.00604
       trunk        74    0.98446      1.106     0.195    0.42271
```

We can reject the hypothesis that `mpg` is normally distributed, but we cannot reject that `trunk` is normally distributed.

The values reported under W and W' are the Shapiro–Wilk and Shapiro–Francia test statistics. The tests also report V and V', which are more appealing indexes for departure from normality. The median values of V and V' are 1 for samples from normal populations. Large values indicate nonnormality. The 95% critical values of V (V'), which depend on the sample size, are between 1.2 and 2.4 (2.0 and 2.8); see Royston (1991d). There is no more information in V (V') than in W (W')—one is just the transform of the other.

◁

Example 2

We have data on a variable called **studytime**, which we suspect is distributed lognormally:

```
. use http://www.stata-press.com/data/r15/cancer
(Patient Survival in Drug Trial)
. generate lnstudytime = ln(studytime)
. swilk lnstudytime
                   Shapiro-Wilk W test for normal data
    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
 lnstudytime |         48    0.92731      3.311     2.547    0.00543
```

We can reject the lognormal assumption. We do *not* specify the **lnnormal** option when testing for lognormality. The **lnnormal** option is for three-parameter lognormality.

Example 3

Having discovered that ln(**studytime**) is not distributed normally, we now test that ln(**studytime** $-\ k$) is normally distributed, where k is chosen so that the resulting skewness is zero. We obtain the estimate for k from **lnskew0**; see [R] **lnskew0**:

```
. lnskew0 lnstudytimek = studytime, level(95)
       Transform |         k     [95% Conf. Interval]       Skewness
-----------------+--------------------------------------------------
  ln(studytim-k) |  -11.01181    -infinity  -.9477328      -.0000173
. swilk lnstudytimek, lnnormal
             Shapiro-Wilk W test for 3-parameter lognormal data
    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
lnstudytimek |         48    0.97064      1.337     1.261    0.10363
```

We cannot reject the hypothesis that ln(**studytime** $+ 11.01181$) is distributed normally. We do specify the **lnnormal** option when using an estimated value of k.

Stored results

swilk and **sfrancia** store the following in **r()**:

Scalars

r(N)	number of observations	r(W)	W or W'
r(p)	p-value	r(V)	V or V'
r(z)	z statistic		

Methods and formulas

The Shapiro–Wilk test is based on Shapiro and Wilk (1965) with a new approximation accurate for $4 \leq n \leq 2000$ (Royston 1992). The calculations made by **swilk** are based on Royston (1982, 1992, 1993b).

The Shapiro–Francia test (Shapiro and Francia 1972; Royston 1983; Royston 1993a) is an approximate test that is similar to the Shapiro–Wilk test for very large samples.

The relative merits of the Shapiro–Wilk and Shapiro–Francia tests the versus skewness and kurtosis test have been a subject of debate. The interested reader is directed to the articles in the *Stata Technical Bulletin*. Our recommendation is to use the Shapiro–Francia test whenever possible, that is, whenever dealing with nonaggregated or ungrouped data (Gould and Rogers 1991; Gould 1992b); see [R] **swilk**. If normality is rejected, use `sktest` to determine the source of the problem. As both D'Agostino, Belanger, and D'Agostino (1990) and Royston (1991c) mention, researchers should also examine the normal quantile plot to determine normality rather than blindly relying on a few test statistics. See the `qnorm` command documented in [R] **diagnostic plots** for more information on normal quantile plots.

Samuel Sanford Shapiro (1930–) earned degrees in statistics and engineering from City College of New York, Columbia, and Rutgers. After employment in the U.S. Army and industry, he joined the faculty at Florida International University in 1972. Shapiro has coauthored various texts in statistics and published several papers on distributional testing and other statistical topics.

Acknowledgment

`swilk` and `sfrancia` were written by Patrick Royston of the MRC Clinical Trials Unit, London and coauthor of the Stata Press book *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*.

References

Brzezinski, M. 2012. The Chen–Shapiro test for normality. *Stata Journal* 12: 368–374.

D'Agostino, R. B., A. J. Belanger, and R. B. D'Agostino, Jr. 1990. A suggestion for using powerful and informative tests of normality. *American Statistician* 44: 316–321.

Genest, C., and G. J. Brackstone. 2010. A conversation with Martin Bradbury Wilk. *Statistical Science* 25: 258–273.

Gould, W. W. 1992a. sg3.7: Final summary of tests of normality. *Stata Technical Bulletin* 5: 10–11. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 114–115. College Station, TX: Stata Press.

———. 1992b. sg11.1: Quantile regression with bootstrapped standard errors. *Stata Technical Bulletin* 9: 19–21. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 137–139. College Station, TX: Stata Press.

Gould, W. W., and W. H. Rogers. 1991. sg3.4: Summary of tests of normality. *Stata Technical Bulletin* 3: 20–23. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 106–110. College Station, TX: Stata Press.

Royston, P. 1982. An extension of Shapiro and Wilks's *W* test for normality to large samples. *Applied Statistics* 31: 115–124.

———. 1983. A simple method for evaluating the Shapiro–Francia *W'* test of non-normality. *Statistician* 32: 297–300.

———. 1991a. sg3.2: Shapiro–Wilk and Shapiro–Francia tests. *Stata Technical Bulletin* 3: 19. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, p. 105. College Station, TX: Stata Press.

———. 1991b. sg3.5: Comment on sg3.4 and an improved D'Agostino test. *Stata Technical Bulletin* 3: 23–24. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 110–112. College Station, TX: Stata Press.

———. 1991c. sg3.6: A response to sg3.3: Comment on tests of normality. *Stata Technical Bulletin* 4: 8–9. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 112–114. College Station, TX: Stata Press.

———. 1991d. Estimating departure from normality. *Statistics in Medicine* 10: 1283–1293.

——. 1992. Approximating the Shapiro–Wilk *W*-test for non-normality. *Statistics and Computing* 2: 117–119.

——. 1993a. A pocket-calculator algorithm for the Shapiro–Francia test for non-normality: An application to medicine. *Statistics in Medicine* 12: 181–184.

——. 1993b. A toolkit for testing for non-normality in complete and censored samples. *Statistician* 42: 37–43.

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

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

Also see

[R] **lnskew0** — Find zero-skewness log or Box–Cox transform

[R] **lv** — Letter-value displays

[R] **sktest** — Skewness and kurtosis test for normality

[R] **diagnostic plots** — Distributional diagnostic plots

[MV] **mvtest normality** — Multivariate normality tests