Title

mvtest normality — Multivariate normality tests

Description

`mvtest normality` performs tests for univariate, bivariate, and multivariate normality.

See [MV] **mvtest** for more multivariate tests.

Quick start

Doornik–Hansen omnibus test for multivariate normality for `v1`, `v2`, `v3`, and `v4`

```
mvtest normality v1 v2 v3 v4
```

Also show Henze–Zirkler's consistent test, Mardia's multivariate kurtosis test, and Mardia's multivariate skewness test

```
mvtest normality v1 v2 v3 v4,                          ///
     stats(dhansen hzirkler kurtosis skewness)
```

Same as above

```
mvtest normality v1 v2 v3 v4, stats(all)
```

Also show Doornik–Hansen test for bivariate normality for each pair of variables

```
mvtest normality v1 v2 v3 v4, stats(all) bivariate
```

As above, but show univariate normality tests from `sktest` instead of the bivariate tests

```
mvtest normality v1 v2 v3 v4, stats(all) univariate
```

Show all multivariate, bivariate, and univariate tests of normality for `v1`, `v2`, `v3`, and `v4`

```
mvtest normality v1 v2 v3 v4, stats(all) bivariate univariate
```

Menu

Statistics > Multivariate analysis > MANOVA, multivariate regression, and related > Multivariate test of means, covariances, and normality

Syntax

`mvtest normality` *varlist* [*if*] [*in*] [*weight*] [, *options*]

options	Description
Options	
`univariate`	display tests for univariate normality (`sktest`)
`bivariate`	display tests for bivariate normality (Doornik–Hansen)
`stats(`*stats*`)`	statistics to be computed

stats	Description
`dhansen`	Doornik–Hansen omnibus test; the default
`hzirkler`	Henze–Zirkler's consistent test
`kurtosis`	Mardia's multivariate kurtosis test
`skewness`	Mardia's multivariate skewness test
`all`	all tests listed here

`bootstrap`, `by`, `jackknife`, `rolling`, and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.

Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.

`aweight`s are not allowed with the `jackknife` prefix; see [R] **jackknife**.

`fweight`s are allowed; see [U] **11.1.6 weight**.

Options

Options

`univariate` specifies that tests for univariate normality be displayed, as obtained from `sktest`; see [R] **sktest**.

`bivariate` specifies that the Doornik–Hansen (2008) test for bivariate normality be displayed for each pair of variables.

`stats(`*stats*`)` specifies one or more test statistics for multivariate normality. Multiple *stats* are separated by white space. The following *stats* are available:

`dhansen` produces the Doornik–Hansen (2008) omnibus test.

`hzirkler` produces Henze–Zirkler's (1990) consistent test.

`kurtosis` produces the test based on Mardia's (1970) measure of multivariate kurtosis.

`skewness` produces the test based on Mardia's (1970) measure of multivariate skewness.

`all` is a convenient shorthand for `stats(dhansen hzirkler kurtosis skewness)`.

Remarks and examples

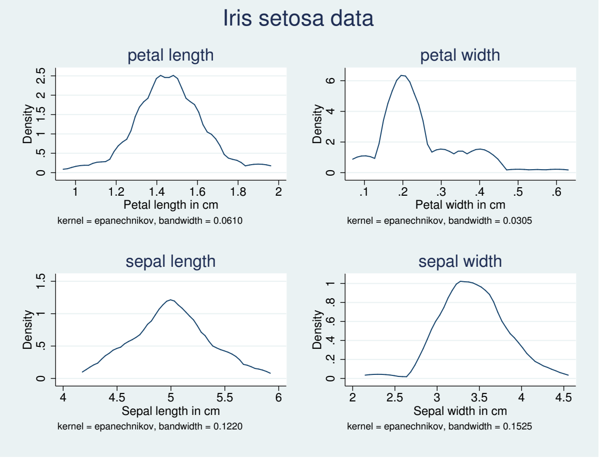
stata.com

Univariate and multivariate tests of normality are provided by the `mvtest normality` command.

▷ Example 1

The classic Fisher iris data from Anderson (1935) consists of four features measured on 50 samples from each of three iris species. The four features are the length and width of the sepal and petal. The three species are *Iris setosa*, *Iris versicolor*, and *Iris virginica*. We hypothesize that these features might be normally distributed within species, though they are likely not normally distributed across species. We will examine the *Iris setosa* data.

```
. use http://www.stata-press.com/data/r15/iris
(Iris data)
. kdensity petlen if iris==1, name(petlen, replace) title(petal length)
. kdensity petwid if iris==1, name(petwid, replace) title(petal width)
. kdensity sepwid if iris==1, name(sepwid, replace) title(sepal width)
. kdensity seplen if iris==1, name(seplen, replace) title(sepal length)
. graph combine petlen petwid seplen sepwid, title("Iris setosa data")
```



We perform all multivariate, univariate, and bivariate tests of normality.

```
. mvtest norm pet* sep* if iris==1, bivariate univariate stats(all)
Test for univariate normality
```

Variable	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
petlen	0.7403	0.1447	2.36	0.3074
petwid	0.0010	0.0442	12.03	0.0024
seplen	0.7084	0.8157	0.19	0.9075
sepwid	0.8978	0.1627	2.07	0.3553

```
Doornik-Hansen test for bivariate normality
```

Pair of variables		chi2	df	Prob>chi2
petlen	petwid	17.47	4	0.0016
	seplen	5.76	4	0.2177
	sepwid	8.50	4	0.0748
petwid	seplen	14.97	4	0.0048
	sepwid	19.15	4	0.0007
seplen	sepwid	5.92	4	0.2049

```
Test for multivariate normality
    Mardia mSkewness =  3.079721    chi2(20) =   27.860   Prob>chi2 =  0.1128
    Mardia mKurtosis =  26.53766     chi2(1) =    1.677   Prob>chi2 =  0.1953
    Henze-Zirkler    =  .9488453     chi2(1) =    2.707   Prob>chi2 =  0.0999
    Doornik-Hansen                   chi2(8) =   24.414   Prob>chi2 =  0.0020
```

From the univariate tests of normality, petwid does not appear to be normally distributed: p-values of 0.0010 for skewness, 0.0442 for kurtosis, and 0.0024 for the joint univariate test. The univariate tests of the other three variables do not lead to a rejection of the null hypothesis of normality.

The bivariate tests of normality show a rejection (at the 5% level) of the null hypothesis of bivariate normality for all pairs of variables that include petwid. Other pairings fail to reject the null hypothesis of bivariate normality.

Of the four multivariate normality tests, only the Doornik–Hansen test rejects the null hypothesis of multivariate normality, p-value of 0.0020.

◁

The Doornik-Hansen (2008) test and Mardia's (1970) test for multivariate kurtosis take computing time roughly proportional to the number of observations. In contrast, the computing time of the test by Henze-Zirkler (1990) and Mardia's (1970) test for multivariate skewness are roughly proportional to the square of the number of observations.

Stored results

`mvtest normality` stores the following in `r()`:

Scalars

`r(p_dh)`	p-value for Doornik–Hansen test (`stats(dhansen)`)
`r(df_dh)`	degrees of freedom of chi2_dh (`stats(dhansen)`)
`r(chi2_dh)`	Doornik–Hansen statistic (`stats(dhansen)`)
`r(rank_hz)`	rank of covariance matrix (`stats(hzirkler)`)
`r(p_hz)`	p-value for two-sided Henze–Zirkler's test (`stats(hzirkler)`)
`r(z_hz)`	normal variate associated with hz (`stats(hzirkler)`)
`r(V_hz)`	expected variance of log(hz) (`stats(hzirkler)`)
`r(E_hz)`	expected value of log(hz) (`stats(hzirkler)`)
`r(hz)`	Henze–Zirkler discrepancy statistic (`stats(hzirkler)`)
`r(rank_mkurt)`	rank of covariance matrix (`stats(kurtosis)`)
`r(p_mkurt)`	p-value for Mardia's multivariate kurtosis test (`stats(kurtosis)`)
`r(z_mkurt)`	normal variate associated with Mardia mKurtosis (`stats(kurtosis)`)
`r(chi2_mkurt)`	chi-squared of Mardia mKurtosis (`stats(kurtosis)`)
`r(mkurt)`	Mardia mKurtosis test statistic (`stats(kurtosis)`)
`r(rank_mskew)`	rank for Mardia mSkewness test (`stats(skewness)`)
`r(p_mskew)`	p-value for Mardia's multivariate skewness test (`stats(skewness)`)
`r(df_mskew)`	degrees of freedom of Mardia mSkewness test (`stats(skewness)`)
`r(chi2_mskew)`	chi-squared of Mardia mSkewness test (`stats(skewness)`)
`r(mskew)`	Mardia mSkewness test statistic (`stats(skewness)`)

Matrices

`r(U_dh)`	matrix with the skewness and kurtosis of orthonormalized variables (used in the Doornik–Hansen test): b1, b2, z(b1), and z(b2) (`stats(dhansen)`)
`r(Btest)`	bivariate test statistics (`bivariate`)
`r(Utest)`	univariate test statistics (`univariate`)

Methods and formulas

There are N independent k-variate observations, $\mathbf{x}_i$, $i = 1, \dots, N$. Let $\mathbf{X}$ denote the $N \times k$ matrix of observations. We wish to test whether these observations are multivariate normal distributed, $\text{MVN}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The sample mean is $\overline{\mathbf{x}} = 1/N \sum_i \mathbf{x}_i$, and the sample covariance matrix is $\mathbf{S} = 1/N \sum (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$.

Methods and formulas are presented under the following headings:

Mardia mSkewness and mKurtosis
Henze–Zirkler
Doornik–Hansen

Mardia mSkewness and mKurtosis

Mardia (1970) defined multivariate skewness, $b_{1,k}$, and kurtosis, $b_{2,k}$, as

$$b_{1,k} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}^3 \qquad \text{and} \qquad b_{2,k} = \frac{1}{N} \sum_{i=1}^{N} g_{ii}^2$$

where $g_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \overline{\mathbf{x}})$. The test statistic

$$z_1 = \frac{(k+1)(N+1)(N+3)}{6\{(N+1)(k+1) - 6\}} b_{1,k}$$

is approximately χ^2 distributed with $k(k+1)(k+2)/6$ degrees of freedom. The test statistic

$$z_2 = \frac{b_{2,k} - k(k+2)}{\sqrt{8k(k+2)/N}}$$

is approximately $N(0,1)$ distributed. Also see Rencher and Christensen (2012, 108); Mardia, Kent, and Bibby (1979, 20–22); and Seber (1984, 148–149).

Henze–Zirkler

The Henze–Zirkler (1990) test, under the assumption that $\mathbf{S}$ is nonsingular, is

$$\begin{aligned} T = {} & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \exp\left\{-\frac{\beta^2}{2}(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right\} \\ & - 2(1+\beta^2)^{-k/2}\sum_{i=1}^{N} \exp\left\{-\frac{\beta^2}{2(1+\beta^2)}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})\right\} \\ & + N(1+2\beta^2)^{-k/2} \end{aligned}$$

where

$$\beta = \frac{1}{\sqrt{2}}\left\{\frac{N(2k+1)}{4}\right\}^{1/(k+4)}$$

As $N \rightarrow \infty$, the first two moments of T are given by

$$E(T) = 1 - (1+2\beta^2)^{-k/2}\left\{1 + \frac{k\beta^2}{1+2\beta^2} + \frac{k(k+2)\beta^4}{2(1+2\beta^2)^2}\right\}$$

$$\begin{aligned} \text{Var}(T) = {} & 2(1+4\beta^2)^{-k/2} + 2(1+2\beta^2)^{-k}\left\{1 + \frac{2k\beta^4}{(1+2\beta^2)^2} + \frac{3k(k+2)\beta^8}{4(1+2\beta^2)^4}\right\} \\ & - 4w^{-k/2}\left\{1 + \frac{3k\beta^4}{2w} + \frac{k(k+2)\beta^8}{2w^2}\right\} \end{aligned}$$

where $w = (1+\beta^2)(1+3\beta^2)$.

Henze–Zirkler suggest obtaining a p-value from the assumption, supported by a series of simulations, that T is approximately lognormal distributed. Thus let $VZ = \ln\left\{1 + \text{Var}(T)/E(T)^2\right\}$ and $EZ = \ln\{E(T)\} - VZ/2$. The transformation $Z = \{\ln(T) - EZ\}/\sqrt{VZ}$. The p-value of Z is computed as $p = 2\Phi(-|Z|)$, where $\Phi()$ is the cumulative normal distribution.

Doornik–Hansen

For the Doornik–Hansen (2008) test, the multivariate observations are transformed, then the univariate skewness and kurtosis for each transformed variable is computed, and then these are combined into an approximate χ^2 statistic.

Let $\mathbf{V}$ be a matrix with ith diagonal element equal to $S_{ii}^{-1/2}$, where S_{ii} is the ith diagonal element of $\mathbf{S}$. $\mathbf{C} = \mathbf{VSV}$ is then the correlation matrix. Let $\mathbf{H}$ be a matrix with columns equal to the eigenvectors of $\mathbf{C}$, and let $\boldsymbol{\Lambda}$ be a diagonal matrix with the corresponding eigenvalues. Let $\breve{\mathbf{X}}$ be the centered version of $\mathbf{X}$, that is, $\overline{\mathbf{x}}$ subtracted from each row. The data are then transformed using $\dot{\mathbf{X}} = \breve{\mathbf{X}}\mathbf{VH}\boldsymbol{\Lambda}^{-1/2}\mathbf{H}'$.

The univariate skewness and kurtosis for each column of $\dot{\mathbf{X}}$ is then computed. The general formula for univariate skewness is $\sqrt{b_1} = m_3/m_2^{3/2}$ and kurtosis is $b_2 = m_4/m_2^2$, where $m_p = 1/N\sum_{i=1}^{N}(x_i - \overline{x})^p$. Let $\dot{x}_i$ denote the ith observation from the selected column of $\dot{\mathbf{X}}$. Because by construction the mean of $\dot{x}$ is zero and the variance m_2 is one, the formulas simplify to $\sqrt{b_1} = m_3$ and $b_2 = m_4$, where $m_p = 1/N\sum_{i=1}^{N}\dot{x}_i^p$.

The univariate skewness, $\sqrt{b_1}$, is transformed into an approximately normal variate, z_1, as in D'Agostino (1970):

$$z_1 = \delta \log\left(y + \sqrt{1+y^2}\right)$$

where

$$y = \left\{\frac{b_1(\omega^2-1)(N+1)(N+3)}{12(N-2)}\right\}^{1/2}$$

$$\delta = \left(\log\sqrt{\omega^2}\right)^{-1/2}$$

$$\omega^2 = -1 + \sqrt{2(\beta-1)}$$

$$\beta = \frac{3(N^2+27N-70)(N+1)(N+3)}{(N-2)(N+5)(N+7)(N+9)}$$

The univariate kurtosis, b_2, is transformed from a gamma variate into a χ^2-variate and then into a standard normal variable, z_2, using the Wilson–Hilferty (1931) transform:

$$z_2 = \sqrt{9\alpha}\left\{\left(\frac{\chi}{2\alpha}\right)^{1/3} - 1 + \frac{1}{9\alpha}\right\}$$

where

$$\chi = 2f(b_2 - 1 - b_1)$$

$$\alpha = a + b_1 c$$

$$f = \frac{(N+5)(N+7)(N^3+37N^2+11N-313)}{12\delta}$$

$$c = \frac{(N-7)(N+5)(N+7)(N^2+2N-5)}{6\delta}$$

$$a = \frac{(N-2)(N+5)(N+7)(N^2+27N-70)}{6\delta}$$

$$\delta = (N-3)(N+1)(N^2+15N-4)$$

The z_1 and z_2 associated with the columns of $\dot{\mathbf{X}}$ are collected into vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The statistic $\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{Z}_2'\mathbf{Z}_2$ is approximately χ^2 distributed with $2k$ degrees of freedom.

Acknowledgment

An earlier implementation of the Doornik and Hansen (2008) test is the `omninorm` package of Baum and Cox (2007).

References

Anderson, E. 1935. The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society* 59: 2–5.

Baum, C. F., and N. J. Cox. 2007. omninorm: Stata module to calculate omnibus test for univariate/multivariate normality. Boston College Department of Economics, Statistical Software Components S417501. https://ideas.repec.org/c/boc/bocode/s417501.html.

D'Agostino, R. B. 1970. Transformation to normality of the null distribution of g_1. *Biometrika* 57: 679–681.

Doornik, J. A., and H. Hansen. 2008. An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* 70: 927–939.

Henze, N., and B. Zirkler. 1990. A class of invariant consistent tests for multivariate normality. *Communications in Statistics—Theory and Methods* 19: 3595–3617.

Mardia, K. V. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57: 519–530.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

Rencher, A. C., and W. F. Christensen. 2012. *Methods of Multivariate Analysis*. 3rd ed. Hoboken, NJ: Wiley.

Seber, G. A. F. 1984. *Multivariate Observations*. New York: Wiley.

Wilson, E. B., and M. M. Hilferty. 1931. The distribution of chi-square. *Proceedings of the National Academy of Sciences* 17: 684–688.

Also see

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

[R] **swilk** — Shapiro–Wilk and Shapiro–Francia tests for normality