ksmirnov — Kolmogorov–Smirnov equality-of-distributions test

Syntax

One-sample Kolmogorov–Smirnov test

`ksmirnov` *varname* = *exp* [*if*] [*in*]

Two-sample Kolmogorov–Smirnov test

`ksmirnov` *varname* [*if*] [*in*] , `by(`*groupvar*`)` [`exact`]

Menu

Statistics > Nonparametric analysis > Tests of hypotheses > Kolmogorov-Smirnov test

Description

`ksmirnov` performs one- and two-sample Kolmogorov–Smirnov tests of the equality of distributions. In the first syntax, *varname* is the variable whose distribution is being tested, and *exp* must evaluate to the corresponding (theoretical) cumulative. In the second syntax, *groupvar* must take on two distinct values. The distribution of *varname* for the first value of *groupvar* is compared with that of the second value.

When testing for normality, please see [R] **sktest** and [R] **swilk**.

Options for two-sample test

Main

`by(`*groupvar*`)` is required. It specifies a binary variable that identifies the two groups.

`exact` specifies that the exact p-value be computed. This may take a long time if $n > 50$.

Remarks and examples

stata.com

▷ Example 1: Two-sample test

Say that we have data on `x` that resulted from two different experiments, labeled as `group==1` and `group==2`. Our data contain

```
. use http://www.stata-press.com/data/r13/ksxmpl
. list

     +-------------+
     | group    x  |
     |-------------|
  1. |     2    2  |
  2. |     1    0  |
  3. |     2    3  |
  4. |     1    4  |
  5. |     1    5  |
     |-------------|
  6. |     2    8  |
  7. |     2   10  |
     +-------------+
```

We wish to use the two-sample Kolmogorov–Smirnov test to determine if there are any differences in the distribution of x for these two groups:

```
. ksmirnov x, by(group)
Two-sample Kolmogorov-Smirnov test for equality of distribution functions
 Smaller group        D       P-value  Corrected
 ----------------------------------------------
 1:                  0.5000    0.424
 2:                 -0.1667    0.909
 Combined K-S:       0.5000    0.785      0.735
```

The first line tests the hypothesis that x for group 1 contains *smaller* values than for group 2. The largest difference between the distribution functions is 0.5. The approximate p-value for this is 0.424, which is not significant.

The second line tests the hypothesis that x for group 1 contains *larger* values than for group 2. The largest difference between the distribution functions in this direction is 0.1667. The approximate p-value for this small difference is 0.909.

Finally, the approximate p-value for the combined test is 0.785, corrected to 0.735. The p-values ksmirnov calculates are based on the asymptotic distributions derived by Smirnov (1933). These approximations are not good for small samples ($n < 50$). They are too conservative—real p-values tend to be substantially smaller. We have also included a less conservative approximation for the nondirectional hypothesis based on an empirical continuity correction—the 0.735 reported in the third column.

That number, too, is only an approximation. An exact value can be calculated using the exact option:

```
. ksmirnov x, by(group) exact
Two-sample Kolmogorov-Smirnov test for equality of distribution functions
 Smaller group        D       P-value      Exact
 ----------------------------------------------
 1:                  0.5000    0.424
 2:                 -0.1667    0.909
 Combined K-S:       0.5000    0.785      0.657
```

◁

▷ Example 2: One-sample test

Let's now test whether x in the example above is distributed normally. Kolmogorov–Smirnov is not a particularly powerful test in testing for normality, and we do not endorse such use of it; see [R] **sktest** and [R] **swilk** for better tests.

In any case, we will test against a normal distribution with the same mean and standard deviation:

```
. summarize x

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           x |          7    4.571429    3.457222          0         10

. ksmirnov x = normal((x-4.571429)/3.457222)

One-sample Kolmogorov-Smirnov test against theoretical distribution
           normal((x-4.571429)/3.457222)

 Smaller group          D        P-value  Corrected
 --------------------------------------------------
 x:                    0.1650    0.683
 Cumulative:          -0.1250    0.803
 Combined K-S:         0.1650    0.991      0.978
```

Because Stata has no way of knowing that we based this calculation on the calculated mean and standard deviation of x, the test statistics will be slightly conservative in addition to being approximations. Nevertheless, they clearly indicate that the data cannot be distinguished from normally distributed data.

◁

Stored results

ksmirnov stores the following in r():

Scalars

r(D_1)	D from line 1	r(D)	combined D
r(p_1)	p-value from line 1	r(p)	combined p-value
r(D_2)	D from line 2	r(p_cor)	corrected combined p-value
r(p_2)	p-value from line 2	r(p_exact)	exact combined p-value

Macros

r(group1)	name of group from line 1	r(group2)	name of group from line 2

Methods and formulas

In general, the Kolmogorov–Smirnov test (Kolmogorov 1933; Smirnov 1933; also see Conover [1999], 428–465) is not very powerful against differences in the tails of distributions. In return for this, it is fairly powerful for alternative hypotheses that involve lumpiness or clustering in the data.

The directional hypotheses are evaluated with the statistics

$$D^+ = \max_x \left\{ F(x) - G(x) \right\}$$

$$D^- = \min_x \left\{ F(x) - G(x) \right\}$$

where $F(x)$ and $G(x)$ are the empirical distribution functions for the sample being compared. The combined statistic is

$$D = \max\left(|D^+|, |D^-| \right)$$

The p-value for this statistic may be obtained by evaluating the asymptotic limiting distribution. Let m be the sample size for the first sample, and let n be the sample size for the second sample. Smirnov (1933) shows that

$$\lim_{m,n\to\infty} \Pr\left\{\sqrt{mn/(m+n)}D_{m,n} \le z\right\} = 1 - 2\sum_{i=1}^{\infty} \left(-1\right)^{i-1} \exp\left(-2i^2z^2\right)$$

The first five terms form the approximation P_a used by Stata. The exact p-value is calculated by a counting algorithm; see Gibbons and Chakraborti (2011, 236–238). A corrected p-value was obtained by modifying the asymptotic p-value by using a numerical approximation technique:

$$Z = \Phi^{-1}(P_a) + 1.04/\min(m,n) + 2.09/\max(m,n) - 1.35/\sqrt{mn/(m+n)}$$
$$p\text{-value} = \Phi(Z)$$

where $\Phi(\cdot)$ is the cumulative normal distribution.

Andrei Nikolayevich Kolmogorov (1903–1987), of Russia, was one of the great mathematicians of the twentieth century, making outstanding contributions in many different branches, including set theory, measure theory, probability and statistics, approximation theory, functional analysis, classical dynamics, and theory of turbulence. He was a faculty member at Moscow State University for more than 60 years.

Nikolai Vasilyevich Smirnov (1900–1966) was a Russian statistician whose work included contributions in nonparametric statistics, order statistics, and goodness of fit. After army service and the study of philosophy and philology, he turned to mathematics and eventually rose to be head of mathematical statistics at the Steklov Mathematical Institute in Moscow.

References

Aivazian, S. A. 1997. Smirnov, Nikolai Vasil'yevich. In *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, ed. N. L. Johnson and S. Kotz, 208–210. New York: Wiley.

Conover, W. J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley.

Gibbons, J. D., and S. Chakraborti. 2011. *Nonparametric Statistical Inference*. 5th ed. Boca Raton, FL: Chapman & Hall/CRC.

Goerg, S. J., and J. Kaiser. 2009. Nonparametric testing of distributions—the Epss–Singleton two-sample test using the empirical characteristic function. *Stata Journal* 9: 454–465.

Jann, B. 2008. Multinomial goodness-of-fit: Large-sample tests with survey design correction and exact tests for small samples. *Stata Journal* 8: 147–169.

Johnson, N. L., and S. Kotz. 1997. Kolmogorov, Andrei Nikolayevich. In *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, ed. N. L. Johnson and S. Kotz, 255–256. New York: Wiley.

Kolmogorov, A. N. 1933. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari* 4: 83–91.

Riffenburgh, R. H. 2012. *Statistics in Medicine*. 3rd ed. San Diego, CA: Academic Press.

Smirnov, N. V. 1933. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University* 2: 3–16.

Also see

[R] **runtest** — Test for random order

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

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