ksmirnov — Kolmogorov–Smirnov equality-of-distributions test

Description

ksmirnov performs one- and two-sample Kolmogorov–Smirnov tests of the equality of distributions. A one-sample test compares the distribution of the tested variable with the specified distribution. A two-sample test tests the equality of the distributions of two samples.

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

Quick start

One-sample test comparing the distribution of v1 with a Student’s t distribution with 5 degrees of freedom

    ksmirnov v1 = t(5,v1)

Two-sample test comparing distributions of v2 in two groups defined by catvar

    ksmirnov v2, by(catvar)

As above, but calculate an exact p-value

    ksmirnov v2, by(catvar) exact

Menu

Statistics > Nonparametric analysis > Tests of hypotheses > Kolmogorov–Smirnov test
Syntax

One-sample Kolmogorov–Smirnov test

\texttt{ksmirnov varname = exp [if] [in]}

Two-sample Kolmogorov–Smirnov test

\texttt{ksmirnov varname [if] [in], by(groupvar) [exact]}

In the first syntax, \textit{varname} is the variable whose distribution is being tested, and \textit{exp} must evaluate to the corresponding (theoretical) cumulative. In the second syntax, \textit{groupvar} must take on two distinct values. The distribution of \textit{varname} for the first value of \textit{groupvar} is compared with that of the second value.

\texttt{collect} is allowed; see [U] 11.1.10 Prefix commands.

Options for two-sample test

- \texttt{by(groupvar)} is required. It specifies a binary variable that identifies the two groups.
- \texttt{exact} specifies that the exact \textit{p}-value be computed.

Remarks and examples

Example 1: Two-sample test

Say that we have data on \textit{x} that resulted from two different experiments, labeled as \texttt{group==1} and \texttt{group==2}. Our data contain

\begin{verbatim}
. use https://www.stata-press.com/data/r17/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 |
  +-----+-----+
\end{verbatim}

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

\begin{verbatim}
. ksmirnov x, by(group)
Two-sample Kolmogorov-Smirnov test for equality of distribution functions
Smaller group | D    | p-value
---------------+------|---------
 1              | 0.5000 | 0.424   
 2              | -0.1667| 0.909   
 Combined K-S   | 0.5000 | 0.785   
\end{verbatim}

stata.com
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 asymptotic $p$-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 asymptotic $p$-value for this small difference is 0.909.

Finally, the approximate asymptotic $p$-value for the combined test is 0.785. The approximate $p$-values ksmirnov calculates are based on the five-term approximation of the asymptotic distributions derived by Smirnov (1933). These approximations are not good for small samples ($n < 50$). They are too conservative.

An exact $p$-value can be calculated using the exact option:

```
. ksmirnov x, by(group) exact
Two-sample Kolmogorov–Smirnov test for equality of distribution functions

<table>
<thead>
<tr>
<th>Smaller group</th>
<th>D</th>
<th>p-value</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5000</td>
<td>0.424</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.1667</td>
<td>0.909</td>
<td></td>
</tr>
<tr>
<td>Combined K-S</td>
<td>0.5000</td>
<td>0.785</td>
<td>0.657</td>
</tr>
</tbody>
</table>
```

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)

<table>
<thead>
<tr>
<th>Smaller group</th>
<th>D</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.1650</td>
<td>0.683</td>
</tr>
<tr>
<td>Cumulative</td>
<td>-0.1250</td>
<td>0.803</td>
</tr>
<tr>
<td>Combined K-S</td>
<td>0.1650</td>
<td>0.991</td>
</tr>
</tbody>
</table>
```

Because Stata has no way of knowing that we based this calculation on the calculated mean and standard deviation of $x$, the $p$-values will be slightly conservative in addition to being approximations. Nevertheless, they clearly indicate that the data cannot be distinguished from normally distributed data.
ksmirnov stores the following in r():

 Scalars
 r(D_1)  D from line 1
 r(p_1)  p-value from line 1
 r(D_2)  D from line 2
 r(p_2)  p-value from line 2
 r(D)   combined D
 r(p)   combined p-value
 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

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

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).
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


Also see

[R] runtest — Test for random order

[R] sktest — Skewness and kurtosis tests for normality

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