spearman — Spearman's and Kendall's correlations

Syntax

Spearman's rank correlation coefficients

spearman [*varlist*] [*if*] [*in*] [, *spearman_options*]

Kendall's rank correlation coefficients

ktau [*varlist*] [*if*] [*in*] [, *ktau_options*]

spearman_options	Description
Main	
stats(*spearman_list*)	list of statistics; select up to three statistics; default is stats(rho)
print(#)	significance level for displaying coefficients
star(#)	significance level for displaying with a star
bonferroni	use Bonferroni-adjusted significance level
sidak	use Šidák-adjusted significance level
pw	calculate all pairwise correlation coefficients by using all available data
matrix	display output in matrix form

ktau_options	Description
Main	
stats(*ktau_list*)	list of statistics; select up to six statistics; default is stats(taua)
print(#)	significance level for displaying coefficients
star(#)	significance level for displaying with a star
bonferroni	use Bonferroni-adjusted significance level
sidak	use Šidák-adjusted significance level
pw	calculate all pairwise correlation coefficients by using all available data
matrix	display output in matrix form

by is allowed with spearman and ktau; see [D] **by**.

where the elements of *spearman_list* may be

rho	correlation coefficient
obs	number of observations
p	significance level

and the elements of *ktau_list* may be

`taua`	correlation coefficient τ_a
`taub`	correlation coefficient τ_b
`score`	score
`se`	standard error of score
`obs`	number of observations
`p`	significance level

Menu

spearman

Statistics > Nonparametric analysis > Tests of hypotheses > Spearman's rank correlation

ktau

Statistics > Nonparametric analysis > Tests of hypotheses > Kendall's rank correlation

Description

`spearman` displays Spearman's rank correlation coefficients for all pairs of variables in *varlist* or, if *varlist* is not specified, for all the variables in the dataset.

`ktau` displays Kendall's rank correlation coefficients between the variables in *varlist* or, if *varlist* is not specified, for all the variables in the dataset. `ktau` is intended for use on small- and moderate-sized datasets; it requires considerable computation time for larger datasets.

Options for spearman

Main

`stats(`*spearman_list*`)` specifies the statistics to be displayed in the matrix of output. `stats(rho)` is the default. Up to three statistics may be specified; `stats(rho obs p)` would display the correlation coefficient, number of observations, and significance level. If *varlist* contains only two variables, all statistics are shown in tabular form, and `stats()`, `print()`, and `star()` have no effect unless the `matrix` option is specified.

`print(`*#*`)` specifies the significance level of correlation coefficients to be printed. Correlation coefficients with larger significance levels are left blank in the matrix. Typing `spearman, print(.10)` would list only those correlation coefficients that are significant at the 10% level or lower.

`star(`*#*`)` specifies the significance level of correlation coefficients to be marked with a star. Typing `spearman, star(.05)` would "star" all correlation coefficients significant at the 5% level or lower.

`bonferroni` makes the Bonferroni adjustment to calculated significance levels. This adjustment affects printed significance levels and the `print()` and `star()` options. Thus `spearman, print(.05) bonferroni` prints coefficients with Bonferroni-adjusted significance levels of 0.05 or less.

`sidak` makes the Šidák adjustment to calculated significance levels. This adjustment affects printed significance levels and the `print()` and `star()` options. Thus `spearman, print(.05) sidak` prints coefficients with Šidák-adjusted significance levels of 0.05 or less.

`pw` specifies that correlations be calculated using pairwise deletion of observations with missing values. By default, `spearman` uses casewise deletion, where observations are ignored if any of the variables in *varlist* are missing.

`matrix` forces `spearman` to display the statistics as a matrix, even if *varlist* contains only two variables. `matrix` is implied if more than two variables are specified.

Options for ktau

Main

`stats(`*ktau_list*`)` specifies the statistics to be displayed in the matrix of output. `stats(taua)` is the default. Up to six statistics may be specified; `stats(taua taub score se obs p)` would display the correlation coefficients τ_a, τ_b, score, standard error of score, number of observations, and significance level. If *varlist* contains only two variables, all statistics are shown in tabular form and `stats()`, `print()`, and `star()` have no effect unless the `matrix` option is specified.

`print(#)` specifies the significance level of correlation coefficients to be printed. Correlation coefficients with larger significance levels are left blank in the matrix. Typing `ktau, print(.10)` would list only those correlation coefficients that are significant at the 10% level or lower.

`star(#)` specifies the significance level of correlation coefficients to be marked with a star. Typing `ktau, star(.05)` would "star" all correlation coefficients significant at the 5% level or lower.

`bonferroni` makes the Bonferroni adjustment to calculated significance levels. This adjustment affects printed significance levels and the `print()` and `star()` options. Thus `ktau, print(.05) bonferroni` prints coefficients with Bonferroni-adjusted significance levels of 0.05 or less.

`sidak` makes the Šidák adjustment to calculated significance levels. This adjustment affects printed significance levels and the `print()` and `star()` options. Thus `ktau, print(.05) sidak` prints coefficients with Šidák-adjusted significance levels of 0.05 or less.

`pw` specifies that correlations be calculated using pairwise deletion of observations with missing values. By default, `ktau` uses casewise deletion, where observations are ignored if any of the variables in *varlist* are missing.

`matrix` forces `ktau` to display the statistics as a matrix, even if *varlist* contains only two variables. `matrix` is implied if more than two variables are specified.

Remarks and examples

stata.com

▷ Example 1

We wish to calculate the correlation coefficients among marriage rate (`mrgrate`), divorce rate (`divorce_rate`), and median age (`medage`) in state data. We can calculate the standard Pearson correlation coefficients and significance by typing

```
.use http://www.stata-press.com/data/r13/states2
(State data)

. pwcorr mrgrate divorce_rate medage, sig

             |  mrgrate divorc~e   medage
-------------+---------------------------
     mrgrate |   1.0000
             |
             |
divorce_rate |   0.7895   1.0000
             |   0.0000
             |
      medage |   0.0011  -0.1526   1.0000
             |   0.9941   0.2900
             |
```

We can calculate Spearman's rank correlation coefficients by typing

```
. spearman mrgrate divorce_rate medage, stats(rho p)
(obs=50)

  +-----------------+
  | Key             |
  |-----------------|
  |   rho           |
  |   Sig. level    |
  +-----------------+

             |  mrgrate divorc~e   medage
-------------+---------------------------
     mrgrate |   1.0000
             |
             |
divorce_rate |   0.6933   1.0000
             |   0.0000
             |
      medage |  -0.4869  -0.2455   1.0000
             |   0.0003   0.0857
             |
```

The large difference in the results is caused by one observation. Nevada's marriage rate is almost 10 times higher than the state with the next-highest marriage rate. An important feature of the Spearman rank correlation coefficient is its reduced sensitivity to extreme values compared with the Pearson correlation coefficient.

We can calculate Kendall's rank correlations by typing

```
. ktau mrgrate divorce_rate medage, stats(taua taub p)
(obs=50)

  +-----------------+
  | Key             |
  |-----------------|
  |   tau_a         |
  |   tau_b         |
  |   Sig. level    |
  +-----------------+

             |    mrgrate    divorc~e      medage
-------------+-------------------------------------
     mrgrate |     0.9829
             |     1.0000
             |
             |
divorce_rate |     0.5110      0.9804
             |     0.5206      1.0000
             |     0.0000
             |
      medage |    -0.3486     -0.1698      0.9845
             |    -0.3544     -0.1728      1.0000
             |     0.0004      0.0828
             |
```

There are tied values for variables `mrgrate`, `divorce_rate`, and `medage`, so tied ranks are used. As a result, $\tau_a < 1$ on the diagonal (see *Methods and formulas* for the definition of τ_a). ◁

❑ Technical note

According to Conover (1999, 323), "Spearman's ρ tends to be larger than Kendall's τ in absolute value. However, as a test of significance, there is no strong reason to prefer one over the other because both will produce nearly identical results in most cases."

❑

▷ Example 2

We illustrate `spearman` and `ktau` with the auto data, which contains some missing values.

```
.use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. spearman mpg rep78

 Number of obs =      69
Spearman's rho =       0.3098

Test of Ho: mpg and rep78 are independent
    Prob > |t| =       0.0096
```

Because we specified two variables, `spearman` displayed the sample size, correlation, and p-value in tabular form. To obtain just the correlation coefficient displayed in matrix form, we type

```
. spearman mpg rep78, stats(rho) matrix
(obs=69)

             |    mpg    rep78
-------------+------------------
         mpg | 1.0000
       rep78 | 0.3098   1.0000
```

The `pw` option instructs `spearman` and `ktau` to use all nonmissing observations between a pair of variables when calculating their correlation coefficient. In the output below, some correlations are based on 74 observations, whereas others are based on 69 because 5 observations contain a missing value for `rep78`.

```
. spearman mpg price rep78, pw stats(rho obs p) star(0.01)

 +-----------------+
 | Key             |
 |-----------------|
 |   rho           |
 |   Number of obs |
 |   Sig. level    |
 +-----------------+

             |    mpg    price    rep78
-------------+---------------------------
         mpg | 1.0000
             |     74
             |
             |
       price |-0.5419*  1.0000
             |     74       74
             | 0.0000
             |
       rep78 | 0.3098*  0.1028   1.0000
             |     69       69       69
             | 0.0096   0.4008
             |
```

Finally, the `bonferroni` and `sidak` options provide adjusted significance levels:

```
. ktau mpg price rep78, stats(taua taub score se p) bonferroni
(obs=69)

  Key

   tau_a
   tau_b
   score
   se of score
   Sig. level

                    mpg      price      rep78

        mpg      0.9471
                 1.0000
              2222.0000
               191.8600


      price     -0.3973     1.0000
                -0.4082     1.0000
              -932.0000  2346.0000
               192.4561   193.0682
                 0.0000

      rep78      0.2076     0.0648     0.7136
                 0.2525     0.0767     1.0000
               487.0000   152.0000  1674.0000
               181.7024   182.2233   172.2161
                 0.0224     1.0000
```

Charles Edward Spearman (1863–1945) was a British psychologist who made contributions to correlation, factor analysis, test reliability, and psychometrics. After several years' military service, he obtained a PhD in experimental psychology at Leipzig and became a professor at University College London, where he sustained a long program of work on the interpretation of intelligence tests. Ironically, the rank correlation version bearing his name is not the formula he advocated.

Maurice George Kendall (1907–1983) was a British statistician who contributed to rank correlation, time series, multivariate analysis, among other topics, and wrote many statistical texts. Most notably, perhaps, his advanced survey of the theory of statistics went through several editions, later ones with Alan Stuart; the baton has since passed to others. Kendall was employed in turn as a government and business statistician, as a professor at the London School of Economics, as a consultant, and as director of the World Fertility Survey. He was knighted in 1974.

Stored results

spearman stores the following in r():

Scalars	
r(N)	number of observations (last variable pair)
r(rho)	ρ (last variable pair)
r(p)	two-sided p-value (last variable pair)
Matrices	
r(Nobs)	number of observations
r(Rho)	ρ
r(P)	two-sided p-value

ktau stores the following in r():

Scalars	
r(N)	number of observations (last variable pair)
r(tau_a)	τ_{a} (last variable pair)
r(tau_b)	τ_{b} (last variable pair)
r(score)	Kendall's score (last variable pair)
r(se_score)	se of score (last variable pair)
r(p)	two-sided p-value (last variable pair)
Matrices	
r(Nobs)	number of observations
r(Tau_a)	τ_{a}
r(Tau_b)	τ_{b}
r(Score)	Kendall's score
r(Se_Score)	standard error of score
r(P)	two-sided p-value

Methods and formulas

Spearman's (1904) rank correlation is calculated as Pearson's correlation computed on the ranks and average ranks (Conover 1999, 314–315). Ranks are as calculated by egen; see [D] **egen**. The significance is calculated using the approximation

$$p = 2 \times \texttt{ttail}(n-2,\, |\widehat{\rho}|\,\sqrt{n-2}\,/\sqrt{1-\widehat{\rho}^2}\,)$$

For any two pairs of ranks (x_i, y_i) and (x_j, y_j) of one variable pair (*varname*$_1$, *varname*$_2$), $1 \leq i, j \leq n$, where n is the number of observations, define them as concordant if

$$(x_i - x_j)(y_i - y_j) > 0$$

and discordant if this product is less than zero.

Kendall's (1938; also see Kendall and Gibbons [1990] or Bland [2000], 222–225) score S is defined as $C - D$, where C (D) is the number of concordant (discordant) pairs. Let $N = n(n-1)/2$ be the total number of pairs, so τ_{a} is given by

$$\tau_{\mathrm{a}} = S/N$$

and τ_{b} is given by

$$\tau_{\mathrm{b}} = \frac{S}{\sqrt{N-U}\sqrt{N-V}}$$

where

$$U = \sum_{i=1}^{N_1} u_i(u_i - 1)/2$$

$$V = \sum_{j=1}^{N_2} v_j(v_j - 1)/2$$

and where N_1 is the number of sets of tied x values, u_i is the number of tied x values in the ith set, N_2 is the number of sets of tied y values, and v_j is the number of tied y values in the jth set. Under the null hypothesis of independence between *varname*$_1$ and *varname*$_2$, the variance of S is exactly (Kendall and Gibbons 1990, 66)

$$\begin{aligned}\mathrm{Var}(S) =& \frac{1}{18}\left\{n(n-1)(2n+5) - \sum_{i=1}^{N_1} u_i(u_i-1)(2u_i+5) - \sum_{j=1}^{N_2} v_j(v_j-1)(2v_j+5)\right\}\\ &+ \frac{1}{9n(n-1)(n-2)}\left\{\sum_{i=1}^{N_1} u_i(u_i-1)(u_i-2)\right\}\left\{\sum_{j=1}^{N_2} v_j(v_j-1)(v_j-2)\right\}\\ &+ \frac{1}{2n(n-1)}\left\{\sum_{i=1}^{N_1} u_i(u_i-1)\right\}\left\{\sum_{j=1}^{N_2} v_j(v_j-1)\right\}\end{aligned}$$

Using a normal approximation with a continuity correction,

$$z = \frac{|S| - 1}{\sqrt{\mathrm{Var}(S)}}$$

For the hypothesis of independence, the statistics S, τ_{a}, and τ_{b} produce equivalent tests and give the same significance.

For Kendall's τ, the normal approximation is surprisingly accurate for sample sizes as small as 8, at least for calculating p-values under the null hypothesis for continuous variables. (See Kendall and Gibbons [1990, chap. 4], who also present some tables for calculating exact p-values for $n < 10$.) For Spearman's ρ, the normal approximation requires larger samples to be valid.

Let v be the number of variables specified so that $k = v(v-1)/2$ correlation coefficients are to be estimated. If `bonferroni` is specified, the adjusted significance level is $p' = \min(1, kp)$. If `sidak` is specified, $p' = \min\left\{1, 1-(1-p)^n\right\}$. See *Methods and formulas* in [R] **oneway** for a more complete description of the logic behind these adjustments.

Early work on rank correlation is surveyed by Kruskal (1958).

Acknowledgment

The original version of `ktau` was written by Sean Becketti, a past editor of the *Stata Technical Bulletin* and author of the Stata Press book *Introduction to Time Series Using Stata*.

References

Barnard, G. A. 1997. Kendall, Maurice George. In *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, ed. N. L. Johnson and S. Kotz, 130–132. New York: Wiley.

Bland, M. 2000. *An Introduction to Medical Statistics*. 3rd ed. Oxford: Oxford University Press.

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

David, H. A., and W. A. Fuller. 2007. Sir Maurice Kendall (1907–1983): A centenary appreciation. *American Statistician* 61: 41–46.

Jeffreys, H. 1961. *Theory of Probability*. 3rd ed. Oxford: Oxford University Press.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30: 81–93.

Kendall, M. G., and J. D. Gibbons. 1990. *Rank Correlation Methods*. 5th ed. New York: Oxford University Press.

Kruskal, W. H. 1958. Ordinal measures of association. *Journal of the American Statistical Association* 53: 814–861.

Lovie, P., and A. D. Lovie. 1996. Charles Edward Spearman, F.R.S. (1863–1945). *Notes and Records of the Royal Society of London* 50: 75–88.

Newson, R. B. 2000a. snp15: somersd—Confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin* 55: 47–55. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 312–322. College Station, TX: Stata Press.

——. 2000b. snp15.1: Update to somersd. *Stata Technical Bulletin* 57: 35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 322–323. College Station, TX: Stata Press.

——. 2000c. snp15.2: Update to somersd. *Stata Technical Bulletin* 58: 30. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 323. College Station, TX: Stata Press.

——. 2001. snp15.3: Update to somersd. *Stata Technical Bulletin* 61: 22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 324. College Station, TX: Stata Press.

——. 2003. snp15_4: Software update for somersd. *Stata Journal* 3: 325.

——. 2005. snp15_5: Software update for somersd. *Stata Journal* 5: 470.

——. 2006. Confidence intervals for rank statistics: Percentile slopes, differences, and ratios. *Stata Journal* 6: 497–520.

Seed, P. T. 2001. sg159: Confidence intervals for correlations. *Stata Technical Bulletin* 59: 27–28. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 267–269. College Station, TX: Stata Press.

Spearman, C. E. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15: 72–101.

Wolfe, F. 1997. sg64: pwcorrs: Enhanced correlation display. *Stata Technical Bulletin* 35: 22–25. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 163–167. College Station, TX: Stata Press.

——. 1999. sg64.1: Update to pwcorrs. *Stata Technical Bulletin* 49: 17. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, p. 159. College Station, TX: Stata Press.

Also see

[R] **correlate** — Correlations (covariances) of variables or coefficients

[R] **nptrend** — Test for trend across ordered groups