spearman — Spearman's and Kendall's correlations

Description	Quick start	Menu	Syntax
Options for spearman	Options for ktau	Remarks and examples	Stored results
Methods and formulas	Acknowledgment	References	Also see

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. When there are two variables, an exact p-value can be calculated optionally using permutations.

`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.

Quick start

Spearman's rank correlation coefficient with approximate p-value for `v1` and `v2`

```
spearman v1 v2
```

Same as above, but report an exact p-value calculated using Monte Carlo permutations

```
spearman v1 v2, exact
```

Same as above, but perform 100,000 Monte Carlo permutations rather than the default of 10,000, and set the random-number seed for reproducibility

```
spearman v1 v2, exact(montecarlo, reps(100000) rseed(1234))
```

Display Spearman's rank correlation coefficients in a matrix for all pairs of `v1`, `v2`, and `v3`

```
spearman v1 v2 v3
```

Display p-values as well as correlation coefficients

```
spearman v1 v2 v3, stats(rho p)
```

Same as above, but perform Bonferroni's adjustment to p-values

```
spearman v1 v2 v3, stats(rho p) bonferroni
```

Kendall's rank correlation coefficients, scores, and standard errors of the scores for pairs of `v1`, `v2`, and `v3`

```
ktau v1 v2 v3, stats(taua taub score se)
```

Same as above, but use pairwise instead of casewise deletion

```
ktau v1 v2 v3, stats(taua taub score se) pw
```

Menu

spearman

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

ktau

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

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(#)`	p-value cutoff for displaying coefficients
`star(#)`	p-value cutoff for displaying a star
`bonferroni`	report Bonferroni-adjusted p-values
`sidak`	report Šidák-adjusted p-values
`pw`	calculate each pairwise correlation coefficient using all available data
`matrix`	display output in matrix form
`exact`[`(`*exact_specs*`)`]	report an exact p-value (available only when *varlist* is two variables)

ktau_options	Description
Main	
`stats(`*ktau_list*`)`	list of statistics; select up to six statistics; default is `stats(taua)`
`print(#)`	p-value cutoff for displaying coefficients
`star(#)`	p-value cutoff for displaying a star
`bonferroni`	report Bonferroni-adjusted p-values
`sidak`	report Šidák-adjusted p-values
`pw`	calculate each pairwise correlation coefficient using all available data
`matrix`	display output in matrix form

`by` and `collect` are allowed with `spearman` and `ktau`; see [U] **11.1.10 Prefix commands**.

where the elements of *spearman_list* may be

`rho`	correlation coefficient
`obs`	number of observations
`p`	p-value

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`	p-value

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 p-value. 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 p-value cutoff for correlation coefficients to be printed. Correlation coefficients with larger p-values are left blank in the matrix. Typing `spearman, print(.10)` would display only those correlation coefficients that have p-values less than or equal to 0.10.

`star(#)` specifies the p-value cutoff for correlation coefficients to be marked with a star. Typing `spearman, star(.05)` would star all correlation coefficients that have p-values less than or equal to 0.05.

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

`sidak` makes the Šidák adjustment to p-values. This adjustment affects displayed p-values and the `print()` and `star()` options. Thus, `spearman, print(.05) sidak` prints coefficients with Šidák-adjusted p-values 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.

`exact` and `exact(`*exact_specs*`)` specify that an exact p-value be reported. This option is available only when *varlist* contains only two variables.

`exact` specifies that an exact p-value from a Monte Carlo permutation test be reported. `exact` is a synonym for `exact(montecarlo)`.

`exact(montecarlo` [, *options*] | `enumerate` [, *options*]`)` specifies that an exact p-value be reported in addition to the approximate p-value. Specifying `exact(montecarlo)` does a Monte Carlo permutation test. Specifying `exact(enumerate)` does an enumeration of all possible permutations. Because the number of all possible permutations is typically extremely large, enumeration is feasible only for very small datasets. The number of permutations will be displayed, and you can click on *Break* to stop the computation. The exact p-value is computed by `permute`.

`exact(montecarlo` [, *options*]`)` allows *options* `show`, `reps(#)`, `rseed(#)`, `saving(`*filename* [, *sav_options*]`)`, `level(#)`, `dots(#)`, `nodots`, and `eps(#)`. The `show` option specifies that the table produced by `permute` also be displayed. By default, 10,000 Monte Carlo permutations are done. That is, the default is the same as specifying `exact(montecarlo, reps(10000))`. The default for `dots()` is `dots(100)` when `reps()` is $\geq$ 10,000; otherwise, it is `dots(1)`. See *Options* in [R] **permute**.

`exact(enumerate` [, *options*]`)` allows *options* `show`, `saving(`*filename* [, *sav_options*]`)`, `dots(#)`, `nodots`, and `eps(#)`. The `show` option specifies that the table produced by `permute` also be displayed. The default for `dots()` is `dots(100)`. See *Options* in [R] **permute**.

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 and τ_b, score, standard error of score, number of observations, and p-value. 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 p-value cutoff for correlation coefficients to be printed. Correlation coefficients with larger p-values are left blank in the matrix. Typing `ktau, print(.10)` would display only those correlation coefficients that have p-values less than or equal to 0.10.

`star(#)` specifies the p-value cutoff for correlation coefficients to be marked with a star. Typing `ktau, star(.05)` would star all correlation coefficients that have p-values less than or equal to 0.05.

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

`sidak` makes the Šidák adjustment to p-values. This adjustment affects displayed p-values and the `print()` and `star()` options. Thus, `ktau, print(.05) sidak` prints coefficients with Šidák-adjusted p-values 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

▷ 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 p-values by typing

```
. use https://www.stata-press.com/data/r19/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)
Number of observations = 50

+-----------------+
| Key             |
|-----------------|
|   rho           |
|   p-value       |
+-----------------+

             |  mrgrate divorc~e   medage
-------------+---------------------------
     mrgrate |   1.0000 
             |        .
             |
divorce_rate |   0.6933   1.0000 
             |   0.0000        .
             |
      medage |  -0.4869  -0.2455   1.0000 
             |   0.0004   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 is its reduced sensitivity to extreme values compared with the Pearson correlation.

We can calculate Kendall's rank correlations by typing

```
. ktau mrgrate divorce_rate medage, stats(taua taub p)
Number of observations = 50

  +-----------------+
  | Key             |
  |-----------------|
  |  tau_a          |
  |  tau_b          |
  |  p-value        |
  +-----------------+

             |  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 average ranks are used for the tied values. As a result, $\tau_a < 1$ on the diagonal (see *Methods and formulas* for the definition of τ_a). ◁

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."

Newson (2000a, 2000b, 2000c, 2001, 2003, 2005, 2006) introduces confidence intervals for Kendall's τ_a. The community-contributed `somersd` command provides these confidence intervals along with additional rank statistics such as Somers' D and Harrell's C and their corresponding confidence intervals.

See Seed (2001) for confidence intervals for Spearman's rank correlation.

▷ Example 2

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

```
. use https://www.stata-press.com/data/r19/auto
(1978 automobile data)

. spearman mpg rep78

Number of observations =     69
        Spearman's rho = 0.3098

Test of H0: mpg and rep78 are independent
                  Prob = 0.0098
```

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
Number of observations = 69

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

We can specify the `pw` option with `spearman` and `ktau` so that all nonmissing observations between a pair of variables when calculating their correlation coefficient are used. 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)
Number of observations:
                             min = 69
                             avg = 71
                             max = 74

  +----------------+
  | Key            |
  |----------------|
  |  rho           |
  |  Number of obs |
  |  p-value       |
  +----------------+

             |      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
             |   0.0098   0.4000        .
             |
```

The bonferroni and sidak options provide adjusted p-values:

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

  Key

   tau_a
   tau_b
   score
   se of score
   p-value

                     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           .
```

◁

▷ Example 3

We continue with the auto data and show an example of spearman with the exact option.

```
. set seed 1234
. spearman mpg gear_ratio if foreign == 1, exact
Permutations (10,000): .........1,000.........2,000.........3,000.........4,000.
> ........5,000.........6,000.........7,000.........8,000.........9,000.........
> 10,000 done
Number of observations =     22
        Spearman's rho = 0.4881
Test of H0: mpg and gear_ratio are independent
                  Prob = 0.0222
            Exact prob = 0.0214 (10,000 Monte Carlo permutations)
```

By default, exact does a Monte Carlo permutation test with 10,000 permutations. Because it is a Monte Carlo test, we set the random-number generator seed before running spearman for reproducibility.

The exact p-value from the Monte Carlo permutation test is 0.0214, which is close to the approximate p-value of 0.0222. Note that the approximate p-value is not based on the normal distribution or the t distribution. It is calculated using a beta distribution fit to the first four moments of the null distribution of Spearman's rank correlation, and these four moments are calculated exactly for any value of N, the number of observations. See *Methods and formulas* below.

Let's increase the number of Monte Carlo permutations to 1,000,000 to see how close the approximate p-value is to the exact p-value. We specify `dots(10000)` to see a dot every 10,000th permutation, rather than the default of every 1,000th permutation. This time we set the random-number generator seed using a suboption. The exact p-value is computed by `permute`, and we can see `permute`'s output by specifying the `show` suboption.

```
. spearman mpg gear_ratio if foreign == 1,
> exact(montecarlo, reps(1000000) dots(10000) rseed(1234) show)
Permutations (1,000,000): .........100,000.........200,000.........300,000......
> ...400,000.........500,000.........600,000.........700,000.........800,000....
> .....900,000.........1,000,000 done
Monte Carlo permutation results                  Number of observations =        22
Permutation variable: mpg                        Number of permutations = 1,000,000

-------------------------------------------------------------------------------
             |
             |                                                Monte Carlo error
             |                                              -------------------
           T |    T(obs)        Test       c       n      p  SE(p)   [95% CI(p)]
-------------+-----------------------------------------------------------------
         rho |  .4880642       lower  988530 1000000  .9885  .0001  .9883  .9887
             |                 upper   11510 1000000  .0115  .0001  .0113  .0117
             |             two-sided                  .0230  .0001  .0227  .0233
-------------------------------------------------------------------------------
Notes: For lower one-sided test, c = #{T <= T(obs)} and p = p_lower = c/n.
       For upper one-sided test, c = #{T >= T(obs)} and p = p_upper = c/n.
       For two-sided test, p = 2*min(p_lower, p_upper); SE and CI approximate.
Number of observations =      22
        Spearman's rho = 0.4881
Test of H0: mpg and gear_ratio are independent
                  Prob = 0.0222
            Exact prob = 0.0230 (1,000,000 Monte Carlo permutations)
```

With this increase in the number of permutations, the exact p-value is now calculated as 0.0230. From the table produced by `permute`, we see that this p-value has a confidence interval of $[0.0227, 0.0233]$, accounting for the Monte Carlo error. The approximate p-value of 0.0222 falls outside this confidence interval, but it is still very close. Not bad for $N = 22$.

◁

▷ Example 4

For very small sample sizes, an exact p-value can be computed by enumerating the full permutation distribution. Here is an example using the auto data with sample size $N = 11$.

```
. spearman mpg gear_ratio if foreign == 1 & mpg <= 24,
>         exact(enumerate, dots(10000))
(enumerating all 831,600 possible permutations)
Permutations (831,600): .........100,000.........200,000.........300,000........
> .400,000.........500,000.........600,000.........700,000.........800,000.... d
> one
Number of observations =      11
        Spearman's rho = 0.5636
Test of H0: mpg and gear_ratio are independent
                  Prob = 0.0722
            Exact prob = 0.0747 (enumerated all 831,600 permutations)
```

The exact p-value is 0.0747. The approximate p-value is 0.0722, which is quite close to the exact p-value in this case, even with only 11 observations.

For $N = 11$, the permutation distribution consists of $11! = 39{,}916{,}800$ permutations. However, the output above says that there are only 831,600 permutations. This is because the values of `mpg` in this sample are not all unique.

```
. tabulate mpg if foreign == 1 & mpg <= 24

    Mileage |
      (mpg) |      Freq.     Percent        Cum.
------------+-----------------------------------
         14 |          1        9.09        9.09
         17 |          2       18.18       27.27
         18 |          2       18.18       45.45
         21 |          2       18.18       63.64
         23 |          3       27.27       90.91
         24 |          1        9.09      100.00
------------+-----------------------------------
      Total |         11      100.00
```

The multiplicities in the values of `mpg` yield permutations that give identical results, and we need to enumerate only the permutations that are distinct. From the values of `mpg`, we see that each distinct permutation has a multiplicity of $2!\,2!\,2!\,3! = 48$, and $39{,}916{,}800/48 = 831{,}600$, which reduces considerably the number of permutations that need to be computed.

`permute`, which computes the permutations, permutes the first variable of the two in the `spearman` *varlist* command. So to best exploit this, the variable that produces the most multiplicities in the permutations should be placed first in *varlist* when doing an enumeration.

◁

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)
`r(p_l)`	lower one-sided p-value (last variable pair)
`r(p_u)`	upper one-sided p-value (last variable pair)
`r(p_exact)`	two-sided exact p-value
`r(p_l_exact)`	lower one-sided exact p-value
`r(p_u_exact)`	upper one-sided exact p-value
`r(n_perm)`	number of permutations performed

Macros

`r(exact)`	`"montecarlo"` or `"enumerate"`
`r(rngstate)`	random-number state used for Monte Carlo permutations

Matrices

`r(Nobs)`	number of observations
`r(Rho)`	ρ
`r(P)`	two-sided p-value

If `exact(..., show)` is specified, the stored results from `permute` are returned as well; see *Stored results* in [R] **permute**.

`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)`	standard error 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

Methods and formulas are presented under the following headings:

Spearman's rank correlation
Exact p-values
Kendall's tau

Spearman's rank correlation

Spearman's (1904) rank correlation is calculated as Pearson's correlation computed on the ranks (averaged for ties) (Conover 1999, 314–315). Ranks are as calculated by `egen`; see [D] **egen**.

If x_i and y_i, where $i = 1, 2, \ldots, n$, are the ranks of one variable pair, and n is the number of observations, then Spearman's rank correlation is

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i = (y_i - \overline{y})^2}}$$

where $\overline{x} = (\sum_i x_i)/n$ is the mean of x and $\overline{y}$ is defined similarly.

Under the null hypothesis of independence (or, more generally, exchangeability), the distribution of ρ is given by all the possible permutations of x_i with y_i fixed. For the permutation distribution, an equivalent statistic to ρ is

$$T = \sum_i x_i\, y_i$$

The moments of T for the permutation distribution can be computed exactly. Its mean is

$$E(T) = \frac{1}{n}\{\sum_i x_i\}\{\sum_i y_i\}$$

Assume now that the ranks, x_i and y_i, are adjusted so that their means are zero. With this assumption, we have (Stuart, Ord, and Arnold 1999, eqs. 27.42 and 27.43)

$$E(T^2) = \frac{1}{n-1}\{\sum_i x_i^2\}\{\sum_i y_i^2\}$$

$$\begin{aligned} E(T^3) = {} & \frac{1}{n}\{\sum_i x_i^3\}\{\sum_i y_i^3\} \\ & + \frac{3}{n(n-1)}\{\sum_{i\neq j} x_i^2 x_j\}\{\sum_{i\neq j} y_i^2 y_j\} \\ & + \frac{36}{n(n-1)(n-2)}\{\sum_{i<j<k} x_i x_j x_k\}\{\sum_{i<j<k} y_i y_j y_k\} \end{aligned}$$

$$\begin{aligned} E(T^4) = {} & \frac{1}{n}\{\sum_i x_i^4\}\{\sum_i y_i^4\} \\ & + \frac{4}{n(n-1)}\{\sum_{i\neq j} x_i^3 x_j\}\{\sum_{i\neq j} y_i^3 y_j\} \\ & + \frac{12}{n(n-1)}\{\sum_{i<j} x_i^2 x_j^2\}\{\sum_{i<j} y_i^2 y_j^2\} \\ & + \frac{24}{n(n-1)(n-2)}\{\sum_{i\neq j,k;j<k} x_i^2 x_j x_k\}\{\sum_{i\neq j,k;j<k} y_i^2 y_j y_k\} \\ & + \frac{576}{n(n-1)(n-2)(n-3)}\{\sum_{i<j<k<l} x_i x_j x_k x_l\}\{\sum_{i<j<k<l} y_i y_j y_k y_l\} \end{aligned}$$

Note that Stuart, Ord, and Arnold (1999, eq. 27.43) express $E(T^3)$ and $E(T^4)$ in terms of k statistics, but that formulation is exactly equivalent to the equations given above.

An approximate p-value for Spearman's rank correlation is calculated by fitting a four-parameter beta distribution to the first four moments of T. (See Lord [1965] and Hanson [1991] for a description of the technique of fitting a four-parameter beta distribution to moments of another distribution.) The four-parameter beta distribution with domain $[l, u]$ is

$$f(x) = \frac{(-l+x)^{\alpha-1}(u-x)^{\beta-1}}{(u-l)^{\alpha+\beta-1}B(\alpha,\beta)}$$

where $0 \le l < u \le 1$ and $B(\alpha,\beta)$ is the beta function with $\alpha > 0$ and $\beta > 0$.

The parameters for the beta distribution are calculated as follows. Let m be the mean of T, v its variance, g_3 its skewness, and g_4 its kurtosis (see *Methods and formulas* in [R] **summarize**). Define

$$r = \frac{6(g_4 - g_3^2 - 1)}{6 + 3g_3^2 - 2g_4}$$

and

$$d = 1 - \frac{24(r+1)}{(r+2)(r+3)g_4 - 3(r-6)(r+1)}$$

Let

$$a = \frac{r(1-\sqrt{d})}{2}$$
$$b = \frac{r(1+\sqrt{d})}{2}$$

If $g_3 > 0$, then $\alpha = a$ and $\beta = b$; otherwise, $\alpha = b$ and $\beta = a$. The domain boundaries are given by

$$l = m - \alpha\sqrt{\frac{v(\alpha+\beta+1)}{\alpha\beta}}$$
$$u = m + \beta\sqrt{\frac{v(\alpha+\beta+1)}{\alpha\beta}}$$

To calculate the p-values, we first scale the observed value of T to the domain of beta:

$$s = \frac{T_{\text{obs}} - l}{u - l}$$

Then the lower and upper one-sided p-values are given by

$$p_{\text{lower}} = \texttt{ibeta}(\alpha, \beta, s)$$
$$p_{\text{upper}} = \texttt{ibetatail}(\alpha, \beta, s)$$

where `ibeta` is Stata's two-parameter cumulative beta distribution and `ibetatail` is the function for its upper tail; see [FN] **Statistical functions**. The two-sided p-value is given by

$$p = \min\{1, 2\min(p_{\text{lower}}, p_{\text{upper}})\}$$

Exact p-values

Exact p-values for Spearman's rank correlation are computed by `permute`. For details on the permutation computation, see [R] **permute**.

Kendall's tau

Kendall's τ is calculated in the following manner. For any two pairs of ranks $(x_i,\, y_i)$ and $(x_j,\, y_j)$, $1 \le i,\, j \le n$, 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 [2015], 187–188) 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 $\tau_{\rm a}$ is given by

$$\tau_{\rm a} = \frac{S}{N}$$

and $\tau_{\rm b}$ is given by

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

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, the variance of S is exactly (Kendall and Gibbons 1990, 66)

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

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 p-value.

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 $n < 10$.

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 p-value is $p' = \min(1, kp)$. If `sidak` is specified, $p' = \min\{1, 1-(1-p)^n\}$. 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, author of the Stata Press book *Introduction to Time Series Using Stata, Revised Edition*.

References

Barnard, G. A. 1997. "Kendall, Maurice George". In *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, edited by N. L. Johnson and S. Kotz, 130–132. New York: Wiley. https://doi.org/10.1002/9781118150719.ch37.

Bland, M. 2015. *An Introduction to Medical Statistics*. 4th 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. https://doi.org/10.1198/000313007X169055.

Hanson, B. A. 1991. Method of moments estimates of the four-paramenter beta compound binominal model and the calculation of classification consistency indexes. ACT Research Report Series 91-5, American College Testing Program, Iowa City, IA.

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. https://doi.org/10.2307/2332226.

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. https://doi.org/10.2307/2281954.

Lord, F. M. 1965. A strong true-score theory, with applications. *Psychometrika* 30: 239–270. https://doi.org/10.1007/BF02289490.

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. https://doi.org/10.1098/rsnr.1996.0007.

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. https://doi.org/10.2307/1412159.

Stuart, A., J. K. Ord, and S. F. Arnold. 1999. *Classical Inference and the Linear Model*. Vol. 2A of *Kendall's Advanced Theory of Statistics*, 6th ed. London: Arnold.

Also see

[R] **correlate** — Correlations of variables

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

[R] **permute** — Permutation tests