snp15.y	Confidence intervals for rank order statistics and their differences

Author: Roger Newson, King's College London, UK. Email: roger.newson@kcl.ac.uk Date: 30 May 2005.

Abstract: Rank order or so-called "non-parametric" methods are in fact based on population parameters, which are zero under the null hypothesis. Two of these parameters are Kendall's τ_a and Somers' D, the parameter tested by a Wilcoxon rank-sum test. Confidence limits for these parameters are more informative than P-values alone, for three reasons. First, confidence intervals show that a high P-value does not prove a null hypothesis. Second, for continuous data, Kendall's τ_a can often be used to define robust confidence limits for Pearson's correlation by Greiner's relation. Third, we can define confidence limits for differences between two Kendall's τ_as or Somers' Ds, and these are informative, because a larger Kendall's τ_a or Somers' D cannot be secondary to a smaller one. The program `somersd` calculates confidence intervals for Somers' D or Kendall's τ_a, using jackknife variances. There is a choice of transformations, including Fisher's z, Daniels' arcsine, Greiner's ρ, the z-transform of Greiner's ρ, and Harrell's c. A `cluster` option is available. The estimation results are saved as for a model fit, so that differences can be estimated using `lincom`.

Keywords: Somers' D; Kendall's tau; rank correlation; rank-sum test; Wilcoxon test; confidence intervals; non-parametric methods.

Syntax

`somersd` [*varlist*] [*weight*] [`if` *exp*] [`in` *range*] [`,` `cluster(`*varname*`)` `level(`*#*`)` `taua` `tdist` `transf(`*transformation_name*`)` `cimatrix(`*new_matrix*`)`]

where *transformation_name* is one of

`iden | z | asin | rho | zrho | c`

`fweights`, `iweights` and `pweights` are allowed; see [U] **14.1.6 weight**. They are treated as described in **Methods and Formulas** below.

Description

`somersd` calculates the rank order statistics Somers' D (corresponding to rank-sum tests) and Kendall's τ_a, with confidence limits. Somers' D or τ_a is calculated for the first variable of *varlist* as a predictor of each of the other variables in *varlist*, with estimates and jackknife variances and confidence intervals output and saved in `e()` as if for the parameters of a model fit. It is possible to use `lincom` to output confidence limits for differences between the population Somers' D or Kendall's τ_a values.

Options

`cluster(`*varname*`)` specifies the variable which defines sampling clusters. If `cluster` is defined, then the between-cluster Somers' D or τ_a is calculated, and the variances are calculated assuming that the data are sampled from a population of clusters, rather than a population of observations.

`level(`*#*`)` specifies the confidence level, in percent, for confidence intervals of the estimates; see [R] `level`.

`taua` causes `somersd` to calculate Kendall's τ_a. If `taua` is absent, then `somersd` calculates Somers' D.

`tdist` specifies that the estimates are assumed to have a t-distribution with $n-1$ degrees of freedom, where n is the number of clusters if `cluster` is specified, or the number of observations if `cluster` is not specified.

`transf(`*transformation_name*`)` specifies that the estimates are to be transformed, defining estimates for the transformed population value. `iden` (identity or untransformed) is the default. `z` specifies Fisher's z (the hyperbolic arctangent), `asin` specifies Daniels' arcsine, `rho` specifies Greiner's ρ (Pearson correlation estimated using Greiner's relation), `zrho` specifies the z-transform of Greiner's ρ, and `c` specifies Harrell's c. If the first variable of *varlist* is a binary indicator of a disease and the other variables are quantitative predictors for that disease, then Harrell's c is the area under the reciever operating characteristic (ROC) curve.

`cimatrix(`*new_matrix*`)` specifies an output matrix to be created, containing estimates and confidence limits for the untransformed Somers' D, Kendall's τ_a or Greiner's ρ parameters. If `transf()` is specified, then the confidence limits will be asymmetric and based on symmetric confidence limits for the transformed parameters. This option (like `level`) may be used in replay mode as well as in non-replay mode.

If a *varlist* is supplied, then all options are allowed. If not, then `somersd` replays the previous `somersd` estimation (if available), and the only options allowed are `level` and `cimatrix`.

Methods and Formulas

The population value of Kendall's τ_a (Kendall, 1970) is defined as

$$\tau_{XY} = E\left[\mathrm{sign}(X_1 - X_2)\,\mathrm{sign}(Y_1 - Y_2)\right], \tag{1}$$

where (X_1, Y_1) and (X_2, Y_2) are bivariate random variables sampled independently from the same population, and $E[\cdot]$ denotes expectation. The population value of Somers' D (Somers, 1962) is defined as

$$D_{YX} = \frac{\tau_{XY}}{\tau_{XX}}. \tag{2}$$

Therefore, τ_{XY} is the difference between two probabilities, namely the probability that the larger of the two X-values is associated with the larger of the two Y-values and the probability that the larger X-value is associated with the smaller Y-value. D_{YX} is the difference between the two corresponding conditional probabilities, given that the two X-values are not equal. Somers' D is related to Harrell's c index by the formula $D = 2c - 1$ (see Harrell *et al.*, 1982 and Harrell *et al.*, 1996). Kendall's τ_a is the covariance between $\mathrm{sign}(X_1 - X_2)$ and $\mathrm{sign}(Y_1 - Y_2)$, whereas Somers' D is the regression coefficient of $\mathrm{sign}(Y_1 - Y_2)$ with respect to $\mathrm{sign}(X_1 - X_2)$. (The correlation coefficient between $\mathrm{sign}(X_1 - X_2)$ and $\mathrm{sign}(Y_1 - Y_2)$ is known as Kendall's τ_b, and is the geometric mean of D_{YX} and D_{XY}.)

Given a sample of data points (X_i, Y_i), we may estimate and test the population values of Kendall's τ_a and Somers' D by the corresponding sample statistics $\hat{\tau}_{XY}$ and $\hat{D}_{YX}$. These are commonly known as "non-parametric" statistics, even though τ_{XY} and D_{YX} are parameters. The two Wilcoxon rank-sum tests (see [R] `signrank`) both test hypotheses predicting $D_{YX} = 0$. The two-sample rank-sum test represents the case where X is a binary variable indicating membership of one of two sub-populations. If the binary X-variable indicates that a patient has a disease, and the Y-variable is a continuous diagnostic test indicator with high values indicating a high probability that the patient has the disease, then the area A under the receiver operating characteristic (ROC) curve, or sensitivity-specificity curve, is linked to Somers' D by the relation $D_{YX} = 2A - 1$. (See [R] **roc** or Hanley and McNeil, 1982.) The matched-pairs rank-sum test represents the case where there are paired data (W_{i1}, W_{i2}), such that $X_i = \mathrm{sign}(W_{i1} - W_{i2})$, and $Y_i = |W_{i1} - W_{i2}|$. Kendall's τ_a is usually tested on "continuous" data, using `ktau` (see [R] `spearman`).

There are several reasons for preferring confidence intervals to P-values alone:

1. Non-statisticians often quote a "non-significant" result for a "non-parametric" test and argue as if they have "proved" a null hypothesis, when a confidence interval would show a wide range of other hypotheses which *also* fit the data.
2. In the case of continuous bivariate data, there is a correspondence between Kendall's τ_a and the more familiar Pearson's correlation coefficient ρ, known as Greiner's relation (Kendall, 1970). This states that

$$\rho = \sin\left(\frac{\pi}{2}\tau_a\right), \tag{3}$$

and holds if the joint distribution of X and Y is bivariate normal. Under this relation, Kendall's τ_a-values of 0, $\pm\frac{1}{3}$, $\pm\frac{1}{2}$ and ± 1 correspond to Pearson's correlations of 0, $\pm\frac{1}{2}$, $\pm\frac{1}{\sqrt{2}}$ and ± 1, respectively. A similar correspondence is likely to hold in a wider range of continuous bivariate distributions (Kendall, 1949; Newson, 1987).

3. Kendall's τ_a has the desirable property that a larger τ_a cannot be secondary to a smaller τ_a. That is to say, if a positive τ_{XY} is caused entirely by a monotonic positive relationship of both variables with a third variable W, then τ_{WX} and τ_{WY} must both be greater than τ_{XY}. If we can show that $\tau_{XY} - \tau_{WY} > 0$ (or, equivalently, that $D_{XY} - D_{WY} > 0$), then this implies that the correlation between X and Y is not caused entirely by the influence of W.

To understand the third point, assume that trivariate data points (W_i, X_i, Y_i) are sampled independently from a common population, with discrete probability mass function $f_{W,X,Y}(\cdot,\cdot,\cdot)$ and marginal probability mass function $f_{W,X}(\cdot,\cdot)$. Define the conditional expectation

$$Z(w_1, x_1, w_2, x_2) = E\left[\mathrm{sign}(Y_2 - Y_1) | W_1 = w_1, X_1 = x_1, W_2 = w_2, X_2 = x_2\right] \tag{4}$$

for any w_1 and w_2 in the range of W-values and any x_1 and x_2 in the range of X-values. If we state that the positive relationship between X_i and Y_i is caused entirely by a monotonic positive relationship between both variables and W_i, then that is equivalent to stating that

$$Z(w_1, x_1, w_2, x_2) \geq 0 \tag{5}$$

whenever $w_1 \leq w_2$ and $x_2 \leq x_1$. However, the difference between the two τ_a coefficients is

$$\begin{aligned}\tau_{WY} - \tau_{XY} =& 2\sum_{w}\sum_{x_2<x_1} f_{W,X}(w,x_1)\, f_{W,X}(w,x_2)\, Z(w,x_1,w,x_2)\\ &+ 2\sum_{x}\sum_{w_1<w_2} f_{W,X}(w_1,x)\, f_{W,X}(w_2,x)\, Z(w_1,x,w_2,x)\\ &+ 4\sum_{w_1<w_2}\sum_{x_2<x_1} f_{W,X}(w_1,x_1)\, f_{W,X}(w_2,x_2)\, Z(w_1,x_1,w_2,x_2).\end{aligned} \tag{6}$$

This difference must be non-negative whenever the inequality (5) applies. In particular, if the distribution of the W_i and X_i is nearly continuous, then the difference (6) will be dominated by the third term, representing discordant (W_i, X_i)-pairs. The difference between τ_a-values will then be determined by the ordering of the Y-values when the larger of two W-values is associated with the smaller of two X-values.

We now define the formulae for estimating τ_{XY}, D_{YX} and their differences. We assume the general case where the observations are clustered, which becomes the familiar unclustered case when there is one observation per cluster. Suppose there are n clusters, and the hth cluster contains m_h observations. Define w_{hi}, X_{hi} and Y_{hi} to be the importance weight, X-value and Y-value, respectively, for the ith observation of the hth cluster. (Like most estimation commands, `somersd` treats `iweights` and `pweights` as importance weights, and treats `fweights` as if they denoted a number of identical observations.) Define

$$\begin{aligned} v_{hijk} =& \begin{cases} w_{hi}w_{jk}, & h \neq j \\ 0, & h = j \end{cases}\\ t_{hijk}^{(XY)} =& w_{hi}w_{jk}\,\mathrm{sign}(X_{hi} - X_{jk})\,\mathrm{sign}(Y_{hi} - Y_{jk})\end{aligned} \tag{7}$$

(for any two observations). We will use the usual dot-substitution notation to define (for instance)

$$v_{h.j.} = \sum_{i=1}^{m_h}\sum_{k=1}^{m_j} v_{hijk}, \quad t_{h.j.}^{(XY)} = \sum_{i=1}^{m_h}\sum_{k=1}^{m_j} t_{hijk}^{(XY)}, \quad v_{h...} = \sum_{j=1}^{n} v_{h.j.}, \quad t_{h...}^{(XY)} = \sum_{j=1}^{n} t_{h.j.}^{(XY)}, \tag{8}$$

and any other sums over any other indices. Given that the clusters are sampled independently from a common population of clusters, we can define

$$V = E\left[v_{h.j.}\right], \quad T_{XY} = E\left[t_{h.j.}^{(XY)}\right], \tag{9}$$

for all $h \neq j$. (In the terminology of Hoeffding (1948), these quantities are regular functionals of the cluster population distribution, and the expressions inside the square brackets are kernels of these regular functionals.) The quantities we really want to estimate are Kendall's τ_a and Somers' D, defined respectively by

$$\tau_{XY} = T_{XY}/V, \quad D_{YX} = T_{XY}/T_{XX} = \tau_{XY}/\tau_{XX}. \tag{10}$$

(These are equal to the familiar formulae (1) and (2) if each cluster contains one observation with an importance weight of one.) To estimate these, we use the jackknife method of Arvesen (1969) on the regular functionals (9) and use appropriate Taylor polynomials. The functionals V and T_{XY} are estimated by the Hoeffding (1948) U-statistics

$$\hat{V} = \frac{v_{....}}{n(n-1)}, \quad \hat{T}_{XY} = \frac{t_{....}^{(XY)}}{n(n-1)}, \tag{11}$$

and the respective jackknife pseudovalues corresponding to the hth cluster are given by

$$\begin{aligned}\psi_h^{(V)} =& (n-1)^{-1}v_{....} - (n-2)^{-1}\left[v_{....} - 2v_{h...}\right],\\ \psi_h^{(XY)} =& (n-1)^{-1}t_{....}^{(XY)} - (n-2)^{-1}\left[t_{....}^{(XY)} - 2t_{h...}^{(XY)}\right].\end{aligned} \tag{12}$$

`somersd` calculates correlation measures for a single variable X with a set of Y-variates $(Y^{(1)}, \ldots, Y^{(p)})$. It calculates, in the first instance, the covariance matrix for $\hat{V}$, $\hat{T}_{XX}$, and $\hat{T}_{XY^{(i)}}$ for $1 \leq i \leq p$. This is done using the jackknife influence matrix Υ, which has n rows labelled by the cluster subscripts, and $p+2$ columns labelled (in Stata fashion) by the names V, X, and $Y^{(i)}$ for $1 \leq i \leq p$. It is defined by

$$\Upsilon\left[h, V\right] = \psi_h^{(V)} - \hat{V}, \quad \Upsilon\left[h, X\right] = \psi_h^{(XX)} - \hat{T}_{XX}, \quad \Upsilon\left[h, Y^{(i)}\right] = \psi_h^{(XY^{(i)})} - \hat{T}_{XY^{(i)}}. \tag{13}$$

The jackknife covariance matrix is then equal to

$$\hat{C} = \left[n(n-1)\right]^{-1} \Upsilon' \Upsilon. \tag{14}$$

The estimates for Kendall's τ_a and Somers' D, for variables Y and X, are defined by

$$\hat{\tau}_{XY} = \hat{T}_{XY}/\hat{V}, \quad \hat{D}_{YX} = \hat{T}_{XY}/\hat{T}_{XX}, \tag{15}$$

and the covariance matrices are defined using Taylor polynomials. In the case of Somers' D, we define the $p \times (p+2)$ matrix of estimated derivatives $\hat{\Gamma}^{(D)}$, whose rows are labelled by the names $Y^{(1)}, \ldots, Y^{(p)}$, and whose columns are labelled by $V, X, Y^{(1)}, \ldots, Y^{(p)}$. This matrix is defined by

$$\begin{aligned}
\hat{\Gamma}^{(D)}\left[Y^{(i)}, X\right] &= \frac{\partial \hat{D}_{Y^{(i)}X}}{\partial \hat{T}_{XX}} = -\frac{\hat{T}_{XY^{(i)}}}{\hat{T}_{XX}^2}, \\
\hat{\Gamma}^{(D)}\left[Y^{(i)}, Y^{(i)}\right] &= \frac{\partial \hat{D}_{Y^{(i)}X}}{\partial \hat{T}_{XY^{(i)}}} = \frac{1}{\hat{T}_{XX}},
\end{aligned} \tag{16}$$

all other entries being zero. In the case of Kendall's τ_a, we define a $(p+1) \times (p+2)$ matrix of estimated derivatives $\hat{\Gamma}^{(\tau)}$, whose rows are labelled by $X, Y^{(1)}, \ldots, Y^{(p)}$, and whose columns are labelled by $V, X, Y^{(1)}, \ldots, Y^{(p)}$. This matrix is defined by

$$\begin{aligned}
\hat{\Gamma}^{(\tau)}\left[X, V\right] &= \frac{\partial \hat{\tau}_{XX}}{\partial \hat{V}} = -\frac{\hat{T}_{XX}}{\hat{V}^2}, \\
\hat{\Gamma}^{(\tau)}\left[X, X\right] &= \frac{\partial \hat{\tau}_{XX}}{\partial \hat{T}_{XX}} = \frac{1}{\hat{V}}, \\
\hat{\Gamma}^{(\tau)}\left[Y^{(i)}, V\right] &= \frac{\partial \hat{\tau}_{XY^{(i)}}}{\partial \hat{V}} = -\frac{\hat{T}_{XY^{(i)}}}{\hat{V}^2}, \\
\hat{\Gamma}^{(\tau)}\left[Y^{(i)}, Y^{(i)}\right] &= \frac{\partial \hat{\tau}_{XY^{(i)}}}{\partial \hat{T}_{XY^{(i)}}} = \frac{1}{\hat{V}},
\end{aligned} \tag{17}$$

all other entries again being zero. The estimated dispersion matrices of the Somers' D and τ_a estimates are therefore $\hat{C}^{(D)}$ and $\hat{C}^{(\tau)}$, respectively, defined by

$$\hat{C}^{(D)} = \hat{\Gamma}^{(D)}\,\hat{C}\,\hat{\Gamma}^{(D)\prime}, \quad \hat{C}^{(\tau)} = \hat{\Gamma}^{(\tau)}\,\hat{C}\,\hat{\Gamma}^{(\tau)\prime}. \tag{18}$$

The `transf()` option offers a choice of transformations. Since these are available both for Somers' D and for Kendall's τ_a, we will denote the original estimate as θ (which can stand for D or τ) and the transformed estimate as ζ. They are summarized below, together with their derivatives $d\zeta/d\theta$ and their inverses $\theta(\zeta)$.

`transf()`	Transform name	$\zeta(\theta)$	$d\zeta/d\theta$	$\theta(\zeta)$
`iden`	Untransformed	θ	1	ζ
`z`	Fisher's z	$\text{arctanh}(\theta) = \frac{1}{2}\log[(1+\theta)/(1-\theta)]$	$\left(1-\theta^2\right)^{-1}$	$\tanh(\zeta) = [\exp(2\zeta)-1]/[\exp(2\zeta)+1]$
`asin`	Daniels' arcsine	$\arcsin(\theta)$	$\left(1-\theta^2\right)^{-1/2}$	$\sin(\zeta)$
`rho`	Greiner's ρ	$\sin(\frac{\pi}{2}\theta)$	$\frac{\pi}{2}\cos(\frac{\pi}{2}\theta)$	$(2/\pi)\arcsin(\zeta)$
`zrho`	Greiner's ρ (z-transformed)	$\text{arctanh}[\sin(\frac{\pi}{2}\theta)]$	$\frac{\pi}{2}\cos(\frac{\pi}{2}\theta)[1-\sin(\frac{\pi}{2}\theta)^2]^{-1}$	$(2/\pi)\arcsin[\tanh(\zeta)]$
`c`	Harrell's c	$(\theta+1)/2$	$1/2$	$2\zeta-1$

If `transf()` is specified, then `somersd` displays and saves the transformed estimates and their estimated covariance, instead of the untransformed versions. If $\hat{C}^{(\theta)}$ is the covariance matrix for the untransformed estimates given by (18), and $\hat{\Gamma}^{(\zeta)}$ is the diagonal matrix whose diagonal entries are the $d\zeta/d\theta$ estimates specified in the table, then the transformed parameter and its covariance matrix are

$$\hat{\zeta} = \zeta(\hat{\theta}), \quad \hat{C}^{(\zeta)} = \hat{\Gamma}^{(\zeta)}\,\hat{C}^{(\theta)}\,\hat{\Gamma}^{(\zeta)\,\prime}. \tag{19}$$

Fisher's z-transform was originally recommended for the Pearson correlation coefficient by Fisher (1921) (see also Gayen (1951)), but Edwardes (1995) recommended it specifically for Somers' D on the basis of simulation studies. Daniels' arcsine was suggested as a normalizing transform in Daniels and Kendall (1947). If `transf(z)` or `transf(asin)` is specified, then `somersd` prints asymmetric confidence intervals for the untransformed D or τ_a values, calculated from symmetric confidence intervals for the transformed parameters using the inverse function $\theta(\zeta)$. (This feature corresponds to the `eform` option of other estimation commands.) Greiner's ρ (Kendall, 1970) is based on the relation (3), and is designed to estimate the Pearson correlation coefficient corresponding to the measured τ_a. If `transf(zrho)` is specified, `somersd` prints asymmetric confidence intervals for Greiner's ρ, using the inverse z-transform on symmetric confidence intervals for the z-transformed Greiner's ρ. Harrell's c is usually a reparameterization of Somers' D, and is recommended in Harrell *et al.* (1982) and Harrell *et al.* (1996) as a general measure of the predictive power of a prognostic score arising from a medical test.

Example 1

In the `auto` data, we compare US cars with foreign cars regarding weight and fuel efficiency. First, we use `ranksum` to give significance tests without confidence intervals:

```
. ranksum mpg,by(foreign)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
     foreign |      obs    rank sum    expected
-------------+---------------------------------
    Domestic |       52      1688.5        1950
     Foreign |       22      1086.5         825
-------------+---------------------------------
    combined |       74        2775        2775
unadjusted variance     7150.00
adjustment for ties      -36.95
                     ----------
adjusted variance       7113.05
Ho: mpg(foreign==Domestic) = mpg(foreign==Foreign)
             z =  -3.101
    Prob > |z| =   0.0019
. ranksum weight,by(foreign)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
     foreign |      obs    rank sum    expected
-------------+---------------------------------
    Domestic |       52      2379.5        1950
     Foreign |       22       395.5         825
-------------+---------------------------------
    combined |       74        2775        2775
unadjusted variance     7150.00
adjustment for ties       -1.06
                     ----------
adjusted variance       7148.94
Ho: weight(foreign==Domestic) = weight(foreign==Foreign)
             z =   5.080
    Prob > |z| =   0.0000
```

We note that US cars are typically heavier and travel fewer miles per gallon than foreign cars. For confidence intervals, we use `somersd`:

```
. somersd foreign mpg weight
Somers' D with variable: foreign
Transformation: Untransformed
Valid observations: 74
Symmetric 95% CI
------------------------------------------------------------------------------
             |              Jackknife
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   .4571678    .135146     3.38   0.001     .1922866    .7220491
      weight |  -.7508741   .0832485    -9.02   0.000    -.9140383     -.58771
------------------------------------------------------------------------------
```

We see that, given a randomly-chosen foreign car and a randomly-chosen US car, the foreign car is 46% more likely to travel more miles per gallon than the US car than *vice versa*, with confidence limits from 19% to 72% more likely. However, being foreign seems to be more reliable as a negative predictor of weight than as a positive predictor of "fuel efficiency". We can use `lincom` to define confidence limits for the difference:

```
. lincom -weight-mpg
 ( 1) - mpg - weight = 0
------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .2937063   .0884397     3.32   0.001     .1203677    .4670449
------------------------------------------------------------------------------
```

The difference between Somers' D-values is positive. This indicates that, if there are two cars, one heavier and consuming fewer gallons per mile, the other lighter and consuming more gallons per mile, then the second is more likely to be foreign. So maybe 1970s US cars were not as wasteful as some people think, and were, if anything, more fuel-efficient for their weight than non-US cars at the time. Figure 1 illustrates this graphically. Data points are domestic cars ("D") and foreign cars ("F"). A regression analysis could show the same thing, but Somers' D shows it in stronger terms, without contentious assumptions such as linearity. (On the other hand, a regression model is more informative if its assumptions are true, so the two methods are mutually complementary.)

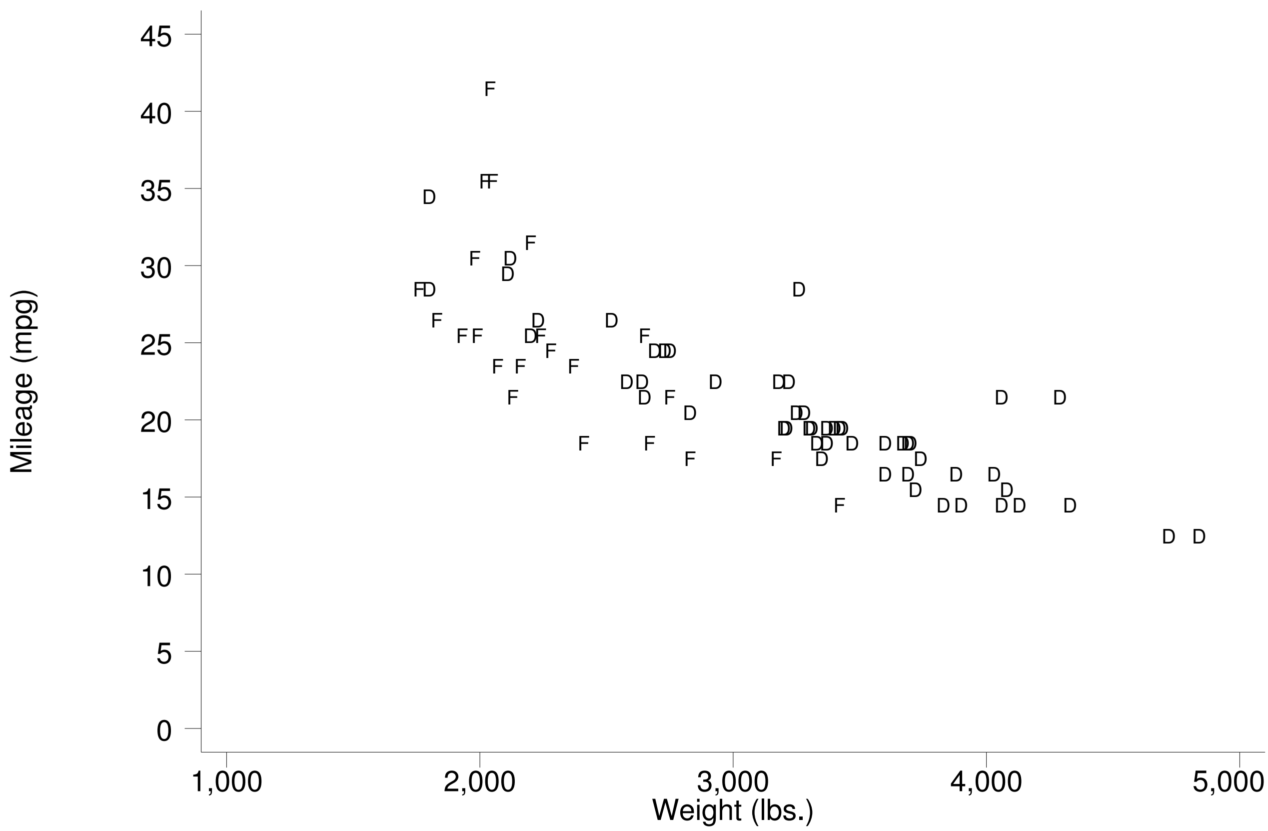

Figure 1. Mileage and weight in US cars (D) and non-US cars (F)

The confidence intervals for such high values of Somers' D would probably be more reliable if we used the z-transform, recommended by Edwardes (1995). The results of this are as follows:

```
. somersd foreign mpg weight,tran(z)
Somers' D with variable: foreign
Transformation: Fisher's z
Valid observations: 74
Symmetric 95% CI for transformed Somers' D
------------------------------------------------------------------------------
             |              Jackknife
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   .4937249   .1708551     2.89   0.004     .1588551    .8285947
      weight |  -.9749561   .1908547    -5.11   0.000    -1.349024   -.6008878
------------------------------------------------------------------------------
Asymmetric 95% CI for untransformed Somers' D
          Somers_D     Minimum     Maximum
   mpg   .45716783   .15753219   .67972072
weight  -.75087413  -.87382282  -.53768098
. lincom -weight-mpg
 ( 1) - mpg - weight = 0
------------------------------------------------------------------------------
     foreign |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .4812312   .1235452     3.90   0.000     .2390871    .7233753
------------------------------------------------------------------------------
```

Note that `somersd` gives not only symmetric confidence limits for the z-transformed Somers' D estimates, but also the more informative asymmetric confidence limits for the untransformed Somers' D estimates (corresponding to the `eform` option). The asymmetric confidence limits for the untransformed estimates are closer to zero than the symmetric confidence limits for the untransformed estimates in the previous output, and are probably more realistic. The output to `lincom` gives confidence limits for the difference between z-transformed Somers' D values. This difference is expressed in z-units, but must, of course, be in the same direction as the difference between untransformed Somers' D values. The conclusions are similar.

Example 2

In this example, we demonstrate Kendall's τ_a by comparing weight (pounds) and displacement (cubic inches) as predictors of fuel efficiency (miles per gallon). We first use `ktau` to carry out significance tests with no confidence limits:

```
. ktau mpg mpg
  Number of obs =      74
Kendall's tau-a =       0.9471
Kendall's tau-b =       1.0000
Kendall's score =    2558
    SE of score =     212.989   (corrected for ties)
Test of Ho: mpg and mpg are independent
     Prob > |z| =       0.0000  (continuity corrected)
. ktau mpg weight
  Number of obs =      74
Kendall's tau-a =      -0.6857
Kendall's tau-b =      -0.7059
Kendall's score =   -1852
    SE of score =     213.605   (corrected for ties)
Test of Ho: mpg and weight are independent
     Prob > |z| =       0.0000  (continuity corrected)
. ktau mpg displ
  Number of obs =      74
Kendall's tau-a =      -0.5942
Kendall's tau-b =      -0.6257
Kendall's score =   -1605
    SE of score =     212.850   (corrected for ties)
Test of Ho: mpg and displ are independent
     Prob > |z| =       0.0000  (continuity corrected)
```

We then use `somersd` (with the `taua` option and the z-transform) to compute the same statistics with confidence limits. Note that `somersd` also outputs the τ_a of `mpg` with `mpg`, which is simply the probability that two independently sampled `mpg`-values are not equal.

```
. somersd mpg weight displ,taua tr(z)
Kendall's tau-a with variable: mpg
Transformation: Fisher's z
Valid observations: 74
Symmetric 95% CI for transformed Kendall's tau-a
------------------------------------------------------------------------------
             |              Jackknife
         mpg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   1.802426   .0748368    24.08   0.000     1.655748    1.949103
      weight |  -.8397412    .084022    -9.99   0.000    -1.004421   -.6750612
       displ |  -.6841711    .093055    -7.35   0.000    -.8665556   -.5017866
------------------------------------------------------------------------------
Asymmetric 95% CI for untransformed Kendall's tau-a
             Tau_a     Minimum     Maximum
   mpg   .94705665   .92964223   .96024957
weight  -.68567197  -.76344472  -.58829928
 displ  -.59422436  -.69961991  -.46352103
```

We can use `lincom` to compare the two predictors and test whether smaller and heavier cars travel fewer miles per gallon than larger and lighter cars. This seems to be the case, as `weight` is a more negative predictor of `mpg` than `displ`:

```
. lincom weight-displ
 ( 1)  weight - displ = 0
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1555701   .0742717    -2.09   0.036    -.3011399   -.0100003
------------------------------------------------------------------------------
```

We demonstrate the `cluster` option using the variable `manuf`, equal to the first word of `make`, to denote manufacturer. This analysis assumes that we are sampling from the population of car manufacturers rather than the population of car models. The results are as follows:

```
. somersd mpg weight displ,taua tr(z) cluster(manuf)
Kendall's tau-a with variable: mpg
Transformation: Fisher's z
Valid observations: 74
Number of clusters: 23
Symmetric 95% CI for transformed Kendall's tau-a
                            (standard errors adjusted for clustering on manuf)
------------------------------------------------------------------------------
             |              Jackknife
         mpg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |    1.83398   .0821029    22.34   0.000     1.673061    1.994898
      weight |  -.8391083   .0917593    -9.14   0.000    -1.018953   -.6592633
       displ |   -.694607   .0976751    -7.11   0.000    -.8860467   -.5031674
------------------------------------------------------------------------------
Asymmetric 95% CI for untransformed Kendall's tau-a
             Tau_a     Minimum     Maximum
   mpg   .95021392   .93195521   .96366535
weight  -.68533644  -.76943983  -.57787293
 displ  -.60093349  -.70943563  -.46460448
. lincom weight-displ
 ( 1)  weight - displ = 0
------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1445012   .0801437    -1.80   0.071      -.30158    .0125775
------------------------------------------------------------------------------
```

Note that, in contrast to the case of most estimation commands, the `cluster` option affects the estimates as well as their standard errors. This is because the clustered estimates are calculated only from between-cluster comparisons, in this case pairs of car models from different manufacturers.

Suppose that we are writing for an audience more familiar with Pearson's correlation than with Kendall's τ_a. To estimate the Pearson correlations corresponding to our τ_a coefficients, we use the `zrho` transform. The results are as follows:

```
. somersd mpg weight displ,taua tr(zrho)
Kendall's tau-a with variable: mpg
Transformation: z-transform of Greiner's rho
Valid observations: 74
Symmetric 95% CI for transformed Greiner's rho
------------------------------------------------------------------------------
             |              Jackknife
         mpg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   3.179521   .1458796    21.80   0.000     2.893602    3.465439
      weight |  -1.378273   .1475561    -9.34   0.000    -1.667478   -1.089069
       displ |  -1.108838    .158893    -6.98   0.000    -1.420262   -.7974132
------------------------------------------------------------------------------
Asymmetric 95% CI for untransformed Greiner's rho
              Rho     Minimum     Maximum
   mpg   .99654393   .99388566   .99804762
weight  -.88056403  -.93121746  -.79653796
 displ  -.80365118  -.88965364  -.66258811
```

The τ_a of -0.59 between displacement and fuel efficiency (from the unclustered output) is seen to correspond to a more impressive Pearson correlation of 0.80. The estimated Greiner's ρ is probably less likely to be oversensitive to outliers than the usual Pearson coefficient.

Saved results

`somersd` saves in `e()`:

Scalars

`e(N)`	number of observations	`e(df_r)`	residual degrees of freedom (if `tdist` present)
`e(N_clust)`	number of clusters		

Macros

`e(cmd)`	`somersd`	`e(param)`	parameter (`somersd` or `taua`)
`e(parmlab)`	parameter label in output	`e(tdist)`	`tdist` if specified
`e(depvar)`	name of X-variable	`e(clustvar)`	name of cluster variable
`e(vcetype)`	covariance estimation method (`Jackknife`)	`e(wtype)`	weight type
`e(wexp)`	weight expression	`e(predict)`	program called by `predict` (set to `somers_p`)
`e(transf)`	transformation specified by `transf`	`e(tranlab)`	transformation label in output

Matrices

`e(b)`	coefficient vector	`e(V)`	variance-covariance matrix of the estimators

Functions

`e(sample)`	marks estimation sample

Note that (confusingly) `e(depvar)` is the X-variable, or predictor variable, in the conventional terminology for defining Somers' D. `somersd` is also different from most estimation commands in that its results are not designed to be used by `predict`. If the user tries to do so, then the program `somers_p` is called, and tells the user that `predict` should not be used after `somersd`.

Historical note

This document is a post-publication update of an article which appeared in the Stata Technical Bulletin (STB) as Newson (2000a). The `somersd` package was later revised in Newson (2000b), Newson (2000c), Newson (2000d), Newson (2001a) and Newson (2001b). An important upgrade (Newson, 2000d) was the addition to the `somersd` package of the program `cendif`, which calculates robust confidence intervals for Hodges-Lehmann median differences, other percentile differences, and percentile ratios. A post-publication update of that STB article is distributed with this document as part of the documentation to the `somersd` package. After 2001, STB was replaced by The Stata Journal (SJ), and all subsequent updates only appeared on SSC and on Roger Newson's homepage at *http://www.kcl-phs.org.uk/rogernewson*, which is accessible from within net-aware Stata. However, Newson (2002) gives a comprehensive review of Somers' D, Kendall's τ_a, median differences, and their estimation in Stata using the `somersd` package.

Acknowledgements

I would like to thank William Gould of Stata Corporation for suggesting the `predict` program `somers_p`, and Nicholas J. Cox of Durham University, UK, for some very helpful discussions on Somers' D and Kendall's τ_a.

References

Arvesen, J. N. 1969. Jackknifing U-statistics. *Annals of Mathematical Statistics* 40: 2076-2100.

Daniels, H. E. and Kendall, M. G. 1947. The Significance of Rank Correlation Where Parental Correlation Exists. *Biometrika* 34: 197-208.

Edwardes, M. D. deB. 1995. A Confidence Interval for Pr($X < Y$) - Pr($X > Y$) Estimated From Simple Cluster Samples. *Biometrics* 51: 571-578.

Fisher, R. A. 1921. On the "Probable Error" of a Coefficient of Correlation deduced from a Small Sample. *Metron* 1(4): 3-32.

Gayen, A. K. 1951. The Frequency Distribution of the Product-Moment Correlation Coefficient in Random Samples of Any Size Drawn from Non-Normal Universes. *Biometrika* 38: 219-247.

Hanley, J. A. and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36.

Harrell, F. E., R. M. Califf, D. B. Pryor, K. L. Lee and R. A. Rosati. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247(18): 2543–2546.

Harrell, F. E., K. L. Lee and D. B. Mark. 1996. Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15: 361–387.

Hoeffding, W. 1948. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 19: 293-325.

Kendall, M. G. 1949. Rank and Product-Moment Correlation. *Biometrika* 36: 177-193.

Kendall, M. G. 1970. *Rank Correlation Methods. 4th edition*. London: Griffin.

Newson, R. B. 1987. An analysis of cinematographic cell division data using U-statistics [D.Phil. dissertation]. Brighton, UK: Sussex University, 301-310.

Newson, R. 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.

Newson, R. 2000b. snp15.1: Update to `somersd`. *Stata Technical Bulletin* 57: 35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 322–323.

Newson, R. 2000c. snp15.2: Update to `somersd`. *Stata Technical Bulletin* 58: 30. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 323.

Newson, R. 2000d. snp16: Robust confidence intervals for median and other percentile differences between groups. *Stata Technical Bulletin* 58: 30–35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 324–331.

Newson, R. 2001a. snp15.3: Update to `somersd`. *Stata Technical Bulletin* 61: 22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 33x.

Newson, R. 2001b. snp16.1: Update to `cendif`. *Stata Technical Bulletin* 61: 22. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 33x.

Newson, R. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal* 2: 45–64. A pre-publication draft can be downloaded from Roger Newson's website at `http://www.kcl-phs.org.uk/rogernewson` using the `net` command in Stata.

Somers, R. H. 1962. A New Asymmetric Measure of Association for Ordinal Variables. *American Sociological Review* 27: 799-811.