The Stata Journal (2025)
24, Number 3, pp. 1–19 DOI: !!

tabagree: Nonparametric measures of agreement and disagreement in paired ordinal data

Milena Falcaro
Queen Mary University of London
London, UK
m.falcaro@qmul.ac.uk

Roger B. Newson
Queen Mary University of London
London, UK
r.newson@qmul.ac.uk

Abstract. In this article, we describe `tabagree`, a new command for assessing the level of agreement and disagreement in paired ordinal data. `tabagree` implements some of the nonparametric measures proposed by Svensson (1993, *Analysis of Systematic and Random Differences Between Paired Ordinal Categorical Data* [Almqvist and Wiksell]) and allows the user to evaluate systematic disagreement separately from random differences. For example, the command can be used in interrater and intrarater reliability studies or in analyses of change.

Keywords: st00!!, tabagree, agreement, ordinal paired data, relative concentration, relative position, relative rank variance

1 Introduction

The need to assess the level of agreement between paired ordinal data arises in many validity and reliability studies. For example, one may be interested in comparing the ratings of two doctors who independently classify the illness severity of a group of patients into five categories (very mild, mild, moderate, severe, and very severe). Disagreement between the two raters may occur because they interpret the categories differently or because one of them tends to systematically rate higher or lower than the other. It may also arise from random error such as, for example, an occasional departure from the measurement protocol or a momentary distraction.

Popular methods in this context are the kappa statistic (κ) and its weighted version (κ_w). The former was initially proposed by Cohen (1960) to adjust the observed agreement by what would be expected by chance alone. Because this treats all disagreements equally, Cohen (1968) later suggested a generalization by introducing the use of weights to account for the different magnitudes of disagreement. Unfortunately, both κ and κ_w have several limitations that may lead to misleading results. Feinstein and Cicchetti (1990) pointed out the paradoxical behavior of κ in certain situations. In particular, they noted that it may be low even in the presence of a high observed agreement. The main problems with κ arise because it depends on the balance and symmetry of the marginal distributions and on the number of categories (Feinstein and Cicchetti 1990; Flight and Julious 2015). κ has also been criticized for not being able to distinguish between different types of disagreement and, in the case of the weighted kappa, for relying on the subjective choice of a set of weights. Several authors have suggested alternative indexes and adjustments to overcome these limitations (see, for example, Gwet [2014] and Klein [2018]). Here we focus on some nonparametric measures proposed by Svensson (1993) for paired ordinal variables.

2 Svensson's method

Let X and Y be two variables measured on n independent statistical units and defined on the same m-category ordinal scale, here encoded by the integers $1, \ldots, m$ for simplicity. The frequency distributions of X and Y can easily be displayed via a contingency table [figure 1(a)]. Svensson (1993) uses a simple alternative representation [figure 1(b)] where the main diagonal of the contingency table is orientated as the main diagonal of a scatterplot, that is, from the lower-left to the upper-right corner. In this way, the table becomes a sort of discrete-version alternative to a scatterplot. Of course, the spacing between categories is artificial, the similarity being the diagonal line of equality. In practice, the output from `tabulate X Y` is in Svensson's representation converted into a contingency table where `X` is the column variable and `Y` is the row variable with categories displayed in descending order (see figure 1).

		Y				
		A	**B**	**C**	**D**	*Total*
X	**A**	30	1	1	2	34
	B	7	10	0	1	18
	C	0	0	25	0	25
	D	0	0	0	33	33
	Total	37	11	26	36	110

(a) tabulate X Y

		X				
		A	**B**	**C**	**D**	*Total*
Y	**D**	2	1	0	33	36
	C	1	0	25	0	26
	B	1	10	0	0	11
	A	30	7	0	0	37
	Total	34	18	25	33	110

(b) Svensson's notation

Figure 1. Example of a contingency table obtained with (a) `tabulate X Y` or (b) Svensson's notation. The categories of X and Y are here labeled as "A", "B", "C", and "D", and alphabetical order is assumed.

Let n_{ij} be the frequency of the pair $(X = i, Y = j)$, where i and $j \in \{1, \ldots, m\}$ and $\sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij} = n$. We also denote with $n_i^{(X)}$ and $n_i^{(Y)}$ the marginal frequencies of the ith category of, respectively, X and Y, and we denote with $C_i^{(X)}$ and $C_i^{(Y)}$ the corresponding cumulative frequencies. For example, for the contingency table in figure 1, the marginal and cumulative frequencies for X are, respectively,

$$\left\{n_1^{(X)}, n_2^{(X)}, n_3^{(X)}, n_4^{(X)}\right\} = \{34, 18, 25, 33\} \quad \text{and}$$

$$\left\{C_1^{(X)}, C_2^{(X)}, C_3^{(X)}, C_4^{(X)}\right\} = \{34\,,\ \underbrace{34+18}_{=52},\ \underbrace{34+18+25}_{=77},\ \underbrace{34+18+25+33}_{=110}\}$$

The percentage agreement (PA) is the proportion of times we observe $X = Y$; that is,

$$\text{PA} = \sum_{i=1}^{m} \frac{n_{ii}}{n}$$

For the contingency table in figure 1(a), we have $\text{PA} = (30 + 10 + 25 + 33)/110 = 89\%$, meaning that the values of X and Y coincide 89 out of 100 times.

The presence of systematic disagreement between X and Y leads to differences in the marginal distributions of the two variables. Svensson (1993) proposed two measures to quantify this type of disagreement: the relative position (RP) and the relative concentration (RC). RP represents the difference between $p_0 = \mathrm{P}(X < Y)$, the probability of X taking lower categories than Y, and $p_1 = \mathrm{P}(X > Y)$, the probability of X taking higher categories than Y. Therefore, it can be defined as

$$\text{RP} = p_0 - p_1$$

where

$$p_0 = \frac{1}{n^2} \sum_{i=1}^{m} \left(n_i^{(Y)} C_{i-1}^{(X)}\right)$$
$$p_1 = \frac{1}{n^2} \sum_{i=1}^{m} \left(n_i^{(X)} C_{i-1}^{(Y)}\right)$$

Possible values for RP range between -1 and 1, with positive values corresponding to situations in which $X < Y$ is more likely to occur than $X > Y$ (higher-scale categories are systematically more frequently used in Y than in X). Equivalently, RP can be written in terms of individual observations as

$$\text{RP} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} \{I(X_k < Y_l) - I(X_k > Y_l)\}$$

where $I(\cdot)$ is an indicator function such that $I(A) = 1$ if the condition A is satisfied and 0 otherwise. Interestingly, as we will show later, RP can also be seen as a special case of Somers's D.

RC measures whether the marginal distribution of Y is systematically more concentrated toward central categories than the marginal distribution of X. It is defined as

$$\mathrm{RC} = \frac{1}{Mn^3} \sum_{i=1}^{m} \left[n_i^{(Y)} C_{i-1}^{(X)} \left\{ n - C_i^{(X)} \right\} - n_i^{(X)} C_{i-1}^{(Y)} \left\{ n - C_i^{(Y)} \right\} \right]$$

where M is a normalizing constant equal to $\min(p_0 - p_0^2, p_1 - p_1^2)$ with $0 < p_0 < 1$ and $0 < p_1 < 1$. RC can take values between -1 and 1 but is not defined if either p_0 or p_1 is equal to 0 or 1. A positive value of RC indicates that Y is more likely than X to have observations in the central part of the scale.

The relative-rank variance (RV) is a rank-based measure of the additional individual variability after adjusting for systematic disagreement and is defined as

$$\mathrm{RV} = \frac{6}{n^3} \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij} \left\{ \overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)} \right\}^2$$

where $\overline{R}_{ij}^{(X)}$ and $\overline{R}_{ij}^{(Y)}$ are the augmented mean ranks for X and Y given by

$$\overline{R}_{ij}^{(X)} = \sum_{k=1}^{i-1} \sum_{l=1}^{m} n_{kl} + \sum_{l=1}^{j-1} n_{il} + \frac{1}{2}(1 + n_{ij})$$

$$\overline{R}_{ij}^{(Y)} = \sum_{k=1}^{m} \sum_{l=1}^{j-1} n_{kl} + \sum_{k=1}^{i-1} n_{kj} + \frac{1}{2}(1 + n_{ij})$$

The higher the value of RV, the more dispersion there is in the observations. Values below 0.1 are generally considered as an indication of negligible individual variation.

Previous simulation studies have shown that RP and RC are approximately normally distributed even for small sample sizes (Kendall 1945). However, Svensson (1993) reported that both exact and asymptotic estimations of the standard errors of RP, RC, and RV are very cumbersome and recommended using bootstrap or jackknife methods (Efron 1981).

The cumulative relative frequencies of the marginal distributions of X and Y can be plotted against each other along with the $(0, 0)$ point to get some sort of relative operating characteristic (ROC) curve; see Svensson (1993) for more details. Note that this use of ROC curves is different from its common application in diagnostic test procedures (for example, Taube [1986]). In this context, the shape of the ROC curve indicates the extent of systematic disagreement. When there is total agreement between X and Y, the ROC curve reduces to the diagonal line from $(0, 0)$ to $(1, 1)$. The curve is S-shaped when there is a systematic difference in concentration, whereas a concave or convex shape is a sign of a systematic shift in position.

The nonparametric measures described in this article have been applied, for example, in studies of change (Svensson 1998; Svensson and Starmark 2002), reliability (Svensson et al. 1996; Allvin et al. 2009), and validity (Lund et al. 2005). For further reading on this topic, see, for example, Svensson and Holm (1994) and Svensson (1997, 1998, 2012).

3 The tabagree command

3.1 Syntax

`tabagree` *var1 var2* [*if*] [*in*] [*weight*] [`,` `table` `display` `label(`*labelname*`)`
`legend` `bsoptions(`*bootstrap_options*`)` `allci` `roc`]

var1 and *var2* can be either string or numeric variables but their values represent only a rank ordering. Value labels attached to *var1* and *var2* are ignored; however, it is possible to use the `label()` option to display value labels rather than numeric codes in the output when `table` or `display` is specified. Swapping the places of *var1* and *var2* (that is, typing `tabagree` *var2 var1*`, ...`) would lead to a change in sign for the RP and RC estimates, but our conclusions would be the same once we account for which variable was first and which was second in the command line.

Only `fweights` (frequency weights) are allowed; see [U] **11.1.6 weight**. Records with zero weight are ignored, as are those in which *var1*, *var2*, or both are missing.

3.2 Options

`table` displays the two-way frequency table of *var1* and *var2*.

`display` shows the contingency table using Svensson's representation, that is, a two-way frequency table where *var2* is the row variable and has its categories displayed in descending order and *var1* is the column variable.

`label(`*labelname*`)` defines the value label for *var1* and *var2* to be used in the result output when `table` or `display` is specified; see [D] **label**.

`legend` displays a legend spelling out the acronyms RP, RC, and RV.

`bsoptions(`*bootstrap_options*`)` instructs Stata to carry out nonparametric bootstrap using the `bootstrap` prefix with *bootstrap_options*. Typing `bsoptions(.)` requests the default bootstrap settings, whereas `bsoptions()` with no argument is equivalent to omitting `bsoptions(`*bootstrap_options*`)` altogether. See [R] **bootstrap**.

`allci` uses the `estat bootstrap` postestimation command to show all available confidence intervals (that is, normal, percentile, bias-corrected, and, if requested, bias-corrected and accelerated confidence intervals). The results are therefore displayed in a table containing the observed value of the statistics, an estimate of their bias, the bootstrap standard errors, and the different confidence intervals. This option is ignored if `bsoptions(`*bootstrap_options*`)` is omitted or specified as `bsoptions()`.

`roc` displays the ROC curve.

3.3 Stored results

`tabagree` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(PA)`	percentage agreement

Macros

`e(cmdname)`	`tabagree`
`e(cmdline)`	command line as typed
`e(properties)`	`b`

Matrices

`e(b)`	vector of estimates

Functions

`e(sample)`	marks estimation sample

If the user requests bootstrapped confidence intervals, then `tabagree` also stores in `e()` additional estimation results stored by `bootstrap`. For example, the estimates of RP, RC, and RV are stored in `e(b)` and the corresponding normal-based confidence intervals in `e(ci_normal)`; see [R] **bootstrap** for more details.

4 Examples

We illustrate the use of `tabagree` by considering a hypothetical interrater agreement study where there are two clinicians (`raterX` and `raterY`) classifying 500 patients into 1 of 4 categories (“A”, “B”, “C”, and “D”, where alphabetical order is assumed). We consider the following three scenarios:

(a)

raterX \ raterY	A	B	C	D	Tot
A	102	9	2	2	115
B	16	99	7	3	125
C	7	35	95	5	142
D	5	4	27	82	118
Tot	130	147	131	92	500

(b)

raterX \ raterY	A	B	C	D	Tot
A	31	49	19	2	101
B	10	94	33	5	142
C	7	30	103	7	147
D	3	7	75	25	110
Tot	51	180	230	39	500

(c)

raterX \ raterY	A	B	C	D	Tot
A	70	63	37	12	182
B	2	52	75	15	144
C	3	3	61	67	134
D	0	1	4	35	40
Tot	75	119	177	129	500

4.1 Scenario (a)

Let’s assume that the dataset is structured as one record per person. If we just want to get the point estimates of Svensson’s measures, we can simply type `tabagree raterX raterY`. If we require confidence intervals for the estimates, we can specify `bsoptions()`. For example, we hereafter specify that we want to perform bootstrap with 200 replications, and we set a random-number seed so that the results can be reproduced. We also specify the `table` option to get the contingency table.

```
. use data_a
. tabagree raterX raterY, table bsoptions(rep(200) seed(123))

Contingency table
           |                     raterY
    raterX |         A          B          C          D |     Total
-----------+--------------------------------------------+----------
         A |       102          9          2          2 |       115
         B |        16         99          7          3 |       125
         C |         7         35         95          5 |       142
         D |         5          4         27         82 |       118
-----------+--------------------------------------------+----------
     Total |       130        147        131         92 |       500

Percentage of agreement = 75.6%

Svensson's measures of agreement and disagreement
(running tabagrsv_rclass on estimation sample)
Bootstrap replications (200): .........10.........20.........30.........40.......
> .50.........60.........70.........80.........90.........100.........110.........
> 120.........130.........140.........150.........160.........170.........180.....
> ....190.........200 done
Bootstrap results                                  Number of obs = 500
                                                   Replications  = 200

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          RP |   -.080956   .0141842    -5.71   0.000    -.1087565   -.0531555
          RC |   .0310298   .0194063     1.60   0.110    -.0070058    .0690654
          RV |   .0408866   .0110364     3.70   0.000     .0192556    .0625176
------------------------------------------------------------------------------
```

In this example, the two raters agreed 75.6% of the time. The disagreement between them was mainly due to differences in how they interpreted the scale categories, `raterX` systematically using higher categories than `raterY` (RP $= -0.08$, 95% confidence interval [CI]: [-0.11 to -0.05]). More specifically, it is 8 percentage points less likely that patients were assigned to higher categories by `raterY` than by `raterX` rather than the opposite. The 95% CI does not contain 0, so the systematic disagreement in position is statistically significant. We also notice that the additional individual variability is negligible (RV < 0.1). In this case, the interrater reliability might be improved by training the raters or making them aware of the bias or both.

4.2 Scenario (b)

Suppose now that the data for this scenario are available only in aggregated form and that in addition to the assessments of `raterX` and `raterY`, the dataset also contains a variable (called `freq`) that indicates the number of records that each observation represents. We can either expand the data before using `tabagree` (that is, type `expand = freq`; this must be followed by `delete if freq==0` if there are empty cells) or simply specify frequency weights in the `tabagree` command line. In this example, we opt for the latter and use the `display`, `bsoptions()`, and `roc` options to get, respectively, the contingency table in Svensson's notation, the bootstrapped confidence intervals, and the

ROC curve. We also specify the **legend** option to get a few extra lines of output that spell out the acronyms used in the results table to denote Svensson's nonparametric statistics.

```
. use data_b, clear

. tabagree raterX raterY [fw=freq], display roc bs(reps(200) seed(1)) legend

Contingency table in Svensson's notation

------------------------------------------------
           |                 raterX
    raterY |      A       B       C       D  Total
-----------+------------------------------------
         D |      2       5       7      25     39
         C |     19      33     103      75    230
         B |     49      94      30       7    180
         A |     31      10       7       3     51
           |
     Total |    101     142     147     110    500
------------------------------------------------

Percentage of agreement = 50.6%

Svensson's measures of agreement and disagreement
 RP: relative position
 RC: relative concentration
 RV: relative rank variance

(running tabagrsv_rclass on estimation sample)

Bootstrap replications (200): .........10.........20.........30.........40........
> .50.........60.........70.........80.........90.........100.........110.........
> 120.........130.........140.........150.........160.........170.........180.....
> ....190.........200 done

Bootstrap results                                          Number of obs = 500
                                                           Replications  = 200

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          RP |   -.010516   .0249305    -0.42   0.673    -.0593789    .0383469
          RC |   .2630043    .024301    10.82   0.000     .2153751    .3106334
          RV |   .1004141   .0172906     5.81   0.000     .0665252     .134303
------------------------------------------------------------------------------
```

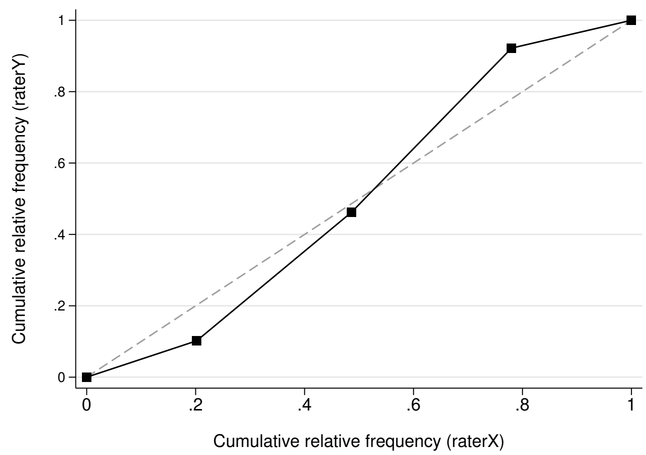

Figure 2. ROC curve created by `tabagree` for scenario (b)

The estimated measure of relative concentration (RC = 0.263, 95% CI: [0.215 to 0.311]) and the S-shaped ROC curve show evidence of systematic differences in concentration. It is more likely that `raterY` uses the central categories more often than `raterX` rather than the opposite.

4.3 Scenario (c)

We now assume that the contingency table is directly entered or imported into Stata and the data look as follows:

```
. use data_c, clear

. list, noobs clean

freq1   freq2   freq3   freq4
   70      63      37      12
    2      52      75      15
    3       3      61      67
    0       1       4      35
```

Before using `tabagree`, we need to convert the dataset into paired observations. This can be done, for example, by using the `reshape` command:

```
. generate raterX=_n
. reshape long freq, i(raterX) j(raterY)
  (output omitted)
. list, noobs clean
   raterX   raterY   freq
        1        1     70
        1        2     63
        1        3     37
        1        4     12
        2        1      2
        2        2     52
        2        3     75
        2        4     15
        3        1      3
        3        2      3
        3        3     61
        3        4     67
        4        1      0
        4        2      1
        4        3      4
        4        4     35
```

The variables `raterX` and `raterY` are now coded with integers from 1 to 4, but we can define a value label that can then be used in the `tabagree` command via the `label()` option. This time, we want `tabagree` to report all available confidence intervals, so we add the `allci` option and increase the number of bootstrap replications to 1,000, with a dot displayed every 100 replications.

```
. label define rlabel 1 "A" 2 "B" 3 "C" 4 "D"
. tabagree raterX raterY [fw=freq], display label(rlabel) roc
> bsoptions(reps(1000) seed(91735) dots(100)) allci

Contingency table in Svensson's notation

-------------------------------------------------
          |                raterX
   raterY |      A       B       C       D  Total
----------+--------------------------------------
        D |     12      15      67      35    129
        C |     37      75      61       4    177
        B |     63      52       3       1    119
        A |     70       2       3       0     75
          |
    Total |    182     144     134      40    500
-------------------------------------------------

Percentage of agreement = 43.6%
```

```
Svensson's measures of agreement and disagreement
(running tabagrsv_rclass on estimation sample)
Bootstrap replications (1,000): .........1,000 done
Bootstrap results                                   Number of obs =    500
                                                    Replications  = 1,000

              Observed                Bootstrap
             coefficient       Bias    std. err.   [95% conf. interval]

        RP      .348256   .0003514   .01930852     .310412       .3861   (N)
                                                   .311964     .386412   (P)
                                                   .311824     .385516  (BC)
        RC   -.02839635  -.0006548   .03393205    -.094902    .0381092   (N)
                                                 -.0958401    .0405957   (P)
                                                 -.0978133    .0398349  (BC)
        RV    .05951395   .0003134   .01202108    .0359531    .0830748   (N)
                                                  .0386144    .0856411   (P)
                                                  .0401599    .0870088  (BC)

Key:  N: Normal
      P: Percentile
     BC: Bias-corrected
```

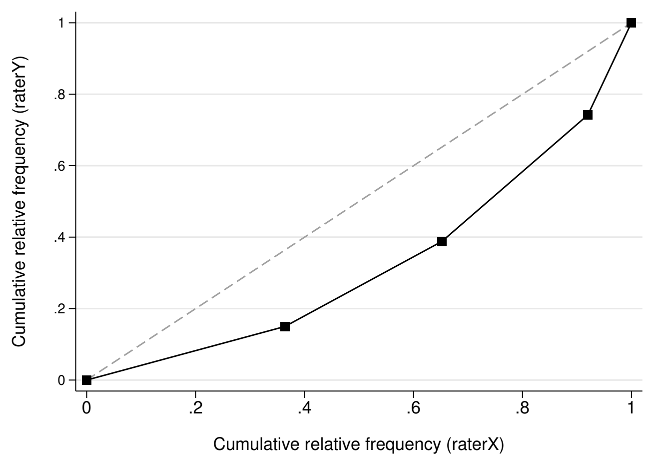

Figure 3. ROC curve created by `tabagree` for scenario (c)

The two marginal distributions differ, which is a sign of systematic discrepancies between the raters. From our table of results, we can infer that the discordance between the raters is mainly due to a systematic disagreement in position (RP = 0.35, 95% CI: [0.31 to 0.39]). Indeed, the ROC curve falls into the right triangle area below the main diagonal, indicating that `raterY` is more likely than `raterX` to assign patients to higher categories. No evidence is found for the presence of significant random differences or systematic disagreement in concentration. Because we specified the `allci` option, the output now contains different types of confidence intervals. Had this option been omitted, Stata would have displayed only the normal-based confidence intervals. Note that the lower and upper normal-based bootstrapped confidence limits may occasionally fall outside the range of possible values (this happens, for example, if the confidence interval for RP contains values below −1 or above 1). Especially in those situations, one may want to estimate the confidence intervals using the bias-corrected or the bias-corrected and accelerated methods because they make direct use of the empirical sampling distribution. As reported in [R] **bootstrap** for the `reps()` option, these methods typically require at least 1,000 replications.

5 A small comparison with kap and somersd using real data

One may wonder how the results from `tabagree` differ from those we can obtain using other nonparametric commands such as, for example, `kap` or the Statistical Software Components package `somersd` (Newson 2002). To answer this, we perform a small comparison using a real data example considered in Agresti (1988) and Holm and Svensson (1991). The data were originally reported in Holmquist, McMahon, and Williams (1967) as part of an interrater reliability study where 7 pathologists had to classify 118 biopsy slides in terms of carcinoma in situ of the uterine cervix. A 5-category ordinal scale was used: 1 = “negative”, 2 = “atypical squamous hyperplasia”, 3 =“carcinoma in situ”, 4 = “squamous carcinoma with early stromal invasion”, and 5 = “invasive carcinoma”. Here we focus on the first two pathologists (labeled as A and B). The dataset contains 1 record for each biopsy slide and 3 variables (`id` = record identifier, `ratingA` = ratings from pathologist A, and `ratingB` = ratings from pathologist B). The 5×5 cross-classification of the ratings is as follows:

```
. use data_pathologistsab, clear
. tabulate ratingA ratingB
           |                         ratingB
   ratingA |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |        22          2          2          0          0 |        26
         2 |         5          7         14          0          0 |        26
         3 |         0          2         36          0          0 |        38
         4 |         0          1         14          7          0 |        22
         5 |         0          0          3          0          3 |         6
-----------+-------------------------------------------------------+----------
     Total |        27         12         69          7          3 |       118
```

The kappa statistics of interrater agreement are then derived as

```
. kap ratingA ratingB
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  63.56%      27.35%     0.4984     0.0482      10.34      0.0000
```

Weighted kappa can be estimated by adding the `wgt()` option with either prerecorded or user-specified weights. For instance, we could specify the prerecorded `w` weights:

```
. kap ratingA ratingB, wgt(w)
Ratings weighted by:
   1.0000   0.7500   0.5000   0.2500   0.0000
   0.7500   1.0000   0.7500   0.5000   0.2500
   0.5000   0.7500   1.0000   0.7500   0.5000
   0.2500   0.5000   0.7500   1.0000   0.7500
   0.0000   0.2500   0.5000   0.7500   1.0000

             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z

  89.62%      70.41%     0.6492     0.0598      10.85      0.0000
```

These estimates of kappa and weighted kappa indicate some disagreement between the pathologists, but they do not provide any indication of why the disagreement arises. Indeed, kappa and weighted kappa do not allow us to distinguish between different sources of disagreement. On the other hand, with Svensson's method, we can get deeper insights and evaluate both the systematic component of interrater differences in terms of RP and RC and the random component as measured by RV:

```
. tabagree ratingA ratingB, display bsoptions(reps(1000) seed(123) dots(50))
> allci roc

Contingency table in Svensson's notation
```

	ratingA					
ratingB	1	2	3	4	5	Total
5	0	0	0	0	3	3
4	0	0	0	7	0	7
3	2	14	36	14	3	69
2	2	7	2	1	0	12
1	22	5	0	0	0	27
Total	26	26	38	22	6	118

```
Percentage of agreement = 63.6%
```

```
Svensson's measures of agreement and disagreement
(running tabagrsv_rclass on estimation sample)
Bootstrap replications (1,000): .........500.........1,000 done
Bootstrap results                                  Number of obs =    118
                                                   Replications  = 1,000
```

	Observed coefficient	Bias	Bootstrap std. err.	[95% conf. interval]		
RP	-.02757828	-.0009337	.03724309	-.1005734	.0454168	(N)
				-.1003304	.048657	(P)
				-.0970267	.0517093	(BC)
RC	.12697865	.0012474	.04882119	.0312909	.2226664	(N)
				.0320597	.2263174	(P)
				.0275384	.2205457	(BC)
RV	.01532289	.0004805	.01135961	-.0069415	.0375873	(N)
				.001888	.0421708	(P)
				.0024321	.0519065	(BC)

```
Key:  N: Normal
      P: Percentile
     BC: Bias-corrected
```

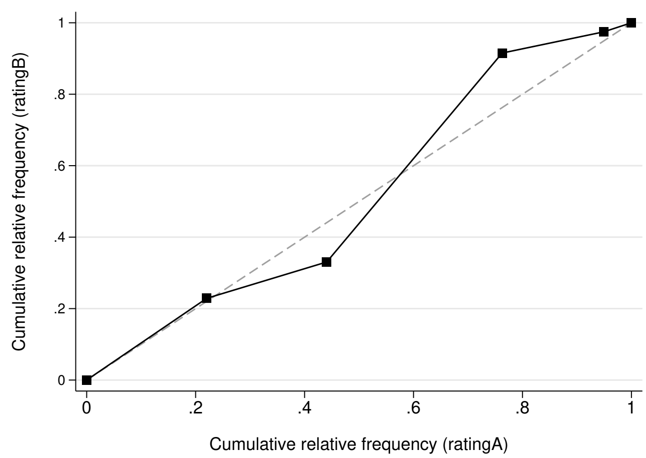

Figure 4. ROC curve created by `tabagree` for the cervical cancer study

These results show that there is a systematic disagreement between the raters. In particular, it is 12.7 (95% bias-corrected CI: [2.8 to 22.1]) percentage points more likely that pathologist B rather than pathologist A uses the central categories more frequently than vice versa. Both pathologists evidently had different opinions about the categories in the middle of the rating scale. Holm and Svensson (1991) argued that "the items in the measuring instrument of the histological classification of carcinoma may be ambiguously described". There is some random variation (RV = 0.015, 95% bias-corrected CI: [0.002 to 0.052]), but it is negligible.

Another nonparametric measure of agreement for paired ordinal variables is Somers's D statistic (Somers 1962), which is implemented in the `somersd` package (Newson 2006). Somers's D has many versions for different variables or sampling schemes, but in our case, it is equal to $P(Y > X) - P(Y < X)$, where X is a random rating by pathologist A and Y is a random rating by pathologist B. These ratings may be for the same subject or for different subjects or for either, depending on the version of Somers's D specified.

To use the `somersd` command, we first need to `reshape` the dataset into a `long` format to have one observation per subject per pathologist and to convert the new within-group identifier (that is, the variable specified in `j()`) into a numeric variable representing the pathologist.

```
. reshape long rating, i(id) j(rater) string
  (output omitted)
. encode rater, generate(pathologist)
```

We can then specify `somersd` with different `funtype()` options to estimate different versions of Somers's D. We first estimate a within-cluster statistic to compare ratings between pathologists within the same subject. This version of Somers's D is the parameter corresponding to a sign test, which is the mean sign of the difference between ratings by the two pathologists for the same subject:

```
. somersd pathologist rating, transf(z) cluster(id) funtype(wcluster)
Within-cluster Somers' D with variable: pathologist
Transformation: Fisher's z
Valid observations: 236
Number of clusters: 118
Symmetric 95% CI for transformed Somers' D
                                  (Std. err. adjusted for 118 clusters in id)
```

pathologist	Coefficient	Jackknife std. err.	z	P>|z|	[95% conf. interval]	
rating	-.0593918	.0557345	-1.07	0.287	-.1686294	.0498458

```
Asymmetric 95% CI for untransformed Somers' D
               Somers_D     Minimum     Maximum
      rating  -.05932203  -.16704898   .04980461
```

The estimated mean sign of the B–A pathologist difference in rating is -0.059 (95% CI: $[-0.169$ to $0.050]$), so it is 5.9 percentage points less likely that pathologist B scores the same subject higher than pathologist A than vice versa.

We then estimate a Von Mises Somers's D, including between-rater comparisons both between subjects and within subjects. This parameter corresponds to a Mann–Whitney or Wilcoxon test comparing all ratings from pathologist B with all ratings from pathologist A, but the confidence limits and p-values are adjusted to allow for clustering by subject. This version of Somers's D is equivalent to Svensson's RP.

```
. somersd pathologist rating, transf(z) cluster(id) funtype(vonmises)
Von Mises Somers' D with variable: pathologist
Transformation: Fisher's z
Valid observations: 236
Number of clusters: 118
Symmetric 95% CI for transformed Somers' D
                                   (Std. err. adjusted for 118 clusters in id)
```

pathologist	Coefficient	Jackknife std. err.	z	P>|z|	[95% conf. interval]	
rating	-.0275853	.0372166	-0.74	0.459	-.1005284	.0453578

```
Asymmetric 95% CI for untransformed Somers' D
               Somers_D      Minimum      Maximum
      rating  -.02757828   -.1001911   .04532675
```

From this, we can conclude that in a random pair of subjects sampled with replacement it is 2.8 (95% CI: [−4.5 to 10.1]) percentage points less likely that pathologist B scores the first subject more highly than pathologist A scores the second subject rather than the opposite.

Note that in both `somersd` commands, we specified the `transf(z)` option, which instructs Stata to use a normalizing Fisher's z (the hyperbolic arctangent) transformation. This computes a symmetric confidence interval for the transformed Somers's D and a back-transformed asymmetric confidence interval for the untransformed Somers's D, ensuring that the lower and upper confidence limits are bounded between −1 and 1.

The `somersd` package can also estimate Kendall's tau-a between $X - Y$ and $X + Y$ (Newson 2002). This tau-a will be positive if absolute X differences tend to be larger than absolute Y differences and will tend to be negative if absolute X differences tend to be smaller than absolute Y differences.

6 Conclusions

Assessing the level of agreement between ordinal paired variables via a single summary index is appealing but usually problematic. The weighted and unweighted kappa statistics, which are the most commonly used measures of agreement in such contexts, have severe limitations because they depend heavily on the marginal distributions and do not distinguish between different sources of disagreement. The weighted kappa offers the advantage of accounting for the ordinal nature of the data but is sensitive to the choice of the weights and, as argued by, for example, Graham and Jackson (1993), it is more a measure of association than of agreement. Somers's D (which includes the sign test statistic and Svensson's RP as special cases) can be used to assess the tendency of one variable to have higher ratings than another, but it does not evaluate the extent of systematic differences in concentration.

In this article, we have described a new command, `tabagree`, that reports alternative rank-invariant measures proposed by Svensson (1993) for the estimation of the systematic (both in position and in concentration) and random components of disagreement. The method has been developed for evaluations between pairs of ordinal variables. Therefore, it is more suited for studies where only one or a few pairwise comparisons are needed. We are not aware of extensions of the method to more than two raters or of a way to incorporate prior knowledge. Nonetheless, when evaluating agreement and disagreement in paired ordinal data, Svensson's measures offer several advantages: they are nonparametric (so they do not rely on strong distributional assumptions); they can be used with small datasets and with zero-frequency cells; and they are easy to interpret.

7 Acknowledgments

This work was supported by Cancer Research UK (grant number: C8162/A27047). We are very grateful to Dr. Tim Morris (MRC Clinical Trials Unit at University College London, UK) for comments on an early draft. We also thank a reviewer and the editor for helping us to improve this manuscript.

8 Programs and supplemental material

To install the software files as they existed at the time of publication of this article, type

```
. net sj 25-3
. net install st00!!      (to install program files, if available)
. net get st00!!          (to install ancillary files, if available)
```

9 References

Agresti, A. 1988. A model for agreement between ratings on an ordinal scale. *Biometrics* 44: 539–548. https://doi.org/10.2307/2531866.

Allvin, R., M. Ehnfors, N. Rawal, E. Svensson, and E. Idvall. 2009. Development of a questionnaire to measure patient-reported postoperative recovery: Content validity and intra-patient reliability. *Journal of Evaluation in Clinical Practice* 15: 411–419. https://doi.org/10.1111/j.1365-2753.2008.01027.x.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46. https://doi.org/10.1177/001316446002000104.

———. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70: 213–220. https://doi.org/10.1037/h0026256.

Efron, B. 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* 68: 589–599. https://doi.org/10.1093/biomet/68.3.589.

Feinstein, A. R., and D. V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43: 543–549. https://doi.org/10.1016/0895-4356(90)90158-l.

Flight, L., and S. A. Julious. 2015. The disagreeable behaviour of the kappa statistic. *Pharmaceutical Statistics* 14: 74–78. https://doi.org/10.1002/pst.1659.

Graham, P., and R. Jackson. 1993. The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology* 46: 1055–1062. https://doi.org/10.1016/0895-4356(93)90173-x.

Gwet, K. L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. 4th ed. Gaithersburg, MD: Advanced Analytics.

Holm, S., and E. Svensson. 1991. Statistical rank methods for ordinal categorical data. Research Report 1991:3, Department of Statistics, University of Göteborg. http://hdl.handle.net/2077/24617.

Holmquist, N. S., C. A. McMahon, and O. D. Williams. 1967. Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology* 84: 334–345.

Kendall, M. G. 1945. The treatment of ties in ranking problems. *Biometrika* 33: 239–251. https://doi.org/10.1093/biomet/33.3.239.

Klein, D. 2018. Implementing a general framework for assessing interrater agreement in Stata. *Stata Journal* 18: 871–901. https://doi.org/10.1177/1536867X1801800408.

Lund, I., T. Lundeberg, L. Sandberg, C. N. Budh, J. Kowalski, and E. Svensson. 2005. Lack of interchangeability between visual analogue and verbal rating pain scales: A cross sectional description of pain etiology groups. *BMC Medical Research Methodology* 5: art. 31. https://doi.org/10.1186/1471-2288-5-31.

Newson, R. B. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal* 2: 45–64. https://doi.org/10.1177/1536867X0200200103.

———. 2006. Confidence intervals for rank statistics: Somers' D and extensions. *Stata Journal* 6: 309–334. https://doi.org/10.1177/1536867X0600600302.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27: 799–811. https://doi.org/10.2307/2090408.

Svensson, E. 1993. *Analysis of Systematic and Random Differences Between Paired Ordinal Categorical Data*. Stockholm: Almqvist and Wiksell.

———. 1997. A coefficient of agreement adjusted for bias in paired ordered categorical data. *Biometrical Journal* 39: 643–657. https://doi.org/10.1002/bimj.4710390602.

———. 1998. Ordinal invariant measures for individual and group changes in ordered categorical data. *Statistics in Medicine* 17: 2923–2936. https://doi.org/10.1002/(SICI)1097-0258(19981230)17:24%3C2923::AID-SIM104%3E3.0.CO;2-%23.

———. 2012. Different ranking approaches defining association and agreement measures of paired ordinal data. *Statistics in Medicine* 31: 3104–3117. https://doi.org/10.1002/sim.5382.

Svensson, E., and S. Holm. 1994. Separation of systematic and random differences in ordinal rating scales. *Statistics in Medicine* 13: 2437–2453. https://doi.org/10.1002/sim.4780132308.

Svensson, E., and J.-E. Starmark. 2002. Evaluation of individual and group changes in social outcome after aneurysmal subarachnoid haemorrhage: A long-term follow-up study. *Journal of Rehabilitation Medicine* 34: 251–259. https://doi.org/10.1080/165019702760390338.

Svensson, E., J.-E. Starmark, S. Ekholm, C. von Essen, and A. Johansson. 1996. Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. *Neurological Research* 18: 487–494. https://doi.org/10.1080/01616412.1996.11740459.

Taube, A. 1986. Sensitivity, specificity and predictive values: A graphical approach. *Statistics in Medicine* 5: 585–591. https://doi.org/10.1002/sim.4780050606.

About the authors

Milena Falcaro is a senior statistician at Queen Mary University of London (UK). Her main research interests are in survival analysis, methods for missing values, and cancer epidemiology.

Roger B. Newson is a senior statistician at Queen Mary University of London (UK), working principally in cancer research. He has written over 120 Statistical Software Components packages (including `somersd`), some of which have been described in detail in articles in the *Stata Journal*.