Title

stata.com

kappa — Interrater agreement

Description

`kap` and `kappa` calculate the kappa-statistic measure of interrater agreement. `kap` calculates the statistic for two unique raters or at least two nonunique raters. `kappa` calculates only the statistic for nonunique raters, but it handles the case where data have been recorded as rating frequencies. `kapwgt` defines weights used by `kap` in measuring the importance of disagreements.

Quick start

Calculate interrater agreement for values `rater1` and `rater2`

```
kap rater1 rater2
```

Add table of assessments

```
kap rater1 rater2, tab
```

As above, and apply frequency weights defined by `wvar`

```
kap rater1 rater2 [fweight=wvar], tab
```

Agreement for values from three nonunique raters stored in `rater1`, `rater2`, and `rater3`

```
kap rater1 rater2 rater3
```

Add values from an additional three raters stored in `rater4`, `rater5`, and `rater6`

```
kap rater1-rater6
```

Use weights $1 - |i - j|/(k - 1)$ to weight disagreements between rater 1 and rater 2

```
kap rater1 rater2, wgt(w)
```

Number of times each subject classified in categories stored in `poor`, `fair`, and `good`

```
kappa poor fair good
```

Menu

kap: two unique raters

Statistics > Epidemiology and related > Other > Interrater agreement, two unique raters

kapwgt

Statistics > Epidemiology and related > Other > Define weights for the above (kap)

kap: nonunique raters

Statistics > Epidemiology and related > Other > Interrater agreement, nonunique raters

kappa

Statistics > Epidemiology and related > Other > Interrater agreement, nonunique raters with frequencies

Syntax

Interrater agreement, two unique raters

`kap` *varname*$_1$ *varname*$_2$ [*if*] [*in*] [*weight*] [`,` *options*]

Weights for weighting disagreements

`kapwgt` *wgtid* [`1 \ # 1` [`\ # # 1 ...`]]

Interrater agreement, nonunique raters, variables record ratings for each rater

`kap` *varname*$_1$ *varname*$_2$ *varname*$_3$ [`...`] [*if*] [*in*] [*weight*]

Interrater agreement, nonunique raters, variables record frequency of ratings

`kappa` *varlist* [*if*] [*in*]

options	Description
Main	
`tab`	display table of assessments
`wgt(`*wgtid*`)`	specify how to weight disagreements; see *Options* for alternatives
`absolute`	treat rating categories as absolute

`collect` is allowed with `kap` and `kappa`; see [U] **11.1.10 Prefix commands**.

`fweights` are allowed; see [U] **11.1.6 weight**.

Options

Main

`tab` displays a tabulation of the assessments by the two raters.

`wgt(`*wgtid*`)` specifies that *wgtid* be used to weight disagreements. You can define your own weights by using `kapwgt`; `wgt()` then specifies the name of the user-defined matrix. For instance, you might define

```
. kapwgt mine 1 \ .8 1 \ 0 .8 1 \ 0 0 .8 1
```

and then

```
. kap rata ratb, wgt(mine)
```

Also, two prerecorded weights are available.

`wgt(w)` specifies weights $1 - |i-j|/(k-1)$, where i and j index the rows and columns of the ratings by the two raters and k is the maximum number of possible ratings.

`wgt(w2)` specifies weights $1 - \{(i-j)/(k-1)\}^2$.

`absolute` is relevant only if `wgt()` is also specified. The `absolute` option modifies how i, j, and k are defined and how corresponding entries are found in a user-defined weighting matrix. When `absolute` is not specified, i and j refer to the row and column index, not to the ratings themselves. Say that the ratings are recorded as $\{0, 1, 1.5, 2\}$. There are four ratings; $k = 4$, and i and j are still 1, 2, 3, and 4 in the formulas above. Index 3, for instance, corresponds to rating $= 1.5$. This system is convenient but can, with some data, lead to difficulties.

When `absolute` is specified, all ratings must be integers, and they must be coded from the set $\{1, 2, 3, \ldots\}$. Not all values need be used; integer values that do not occur are simply assumed to be unobserved.

Remarks and examples

stata.com

Remarks are presented under the following headings:

The kappa-statistic measure of agreement is scaled to be 0 when the amount of agreement is what would be expected to be observed by chance and 1 when there is perfect agreement. For intermediate values, Landis and Koch (1977a, 165) suggest the following interpretations:

below 0.0	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Two raters

▷ Example 1

Consider the classification by two radiologists of 85 xeromammograms as normal, benign disease, suspicion of cancer, or cancer (a subset of the data from Boyd et al. [1982] and discussed in the context of kappa in Altman [1991, 403–405]).

```
. use https://www.stata-press.com/data/r17/rate2
(Altman p. 403)
. tabulate rada radb
```

Radiologist A's assessment	Radiologist B's assessment				
	Normal	Benign	Suspect	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspect	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

Our dataset contains two variables: `rada`, radiologist A's assessment, and `radb`, radiologist B's assessment. Each observation is a patient.

We can obtain the kappa measure of interrater agreement by typing

```
. kap rada radb
```

Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
63.53%	30.82%	0.4728	0.0694	6.81	0.0000

If each radiologist had made his determination randomly (but with probabilities equal to the overall proportions), we would expect the two radiologists to agree on 30.8% of the patients. In fact, they agreed on 63.5% of the patients, or 47.3% of the way between random agreement and perfect agreement. The amount of agreement indicates that we can reject the hypothesis that they are making their determinations randomly.

◁

▷ Example 2: Weighted kappa, prerecorded weight w

There is a difference between two radiologists disagreeing about whether a xeromammogram indicates cancer or the suspicion of cancer and disagreeing about whether it indicates cancer or is normal. The weighted kappa attempts to deal with this. `kap` provides two "prerecorded" weights, `w` and `w2`:

```
. kap rada radb, wgt(w)
Ratings weighted by:
   1.0000   0.6667   0.3333   0.0000
   0.6667   1.0000   0.6667   0.3333
   0.3333   0.6667   1.0000   0.6667
   0.0000   0.3333   0.6667   1.0000
             Expected
Agreement    agreement      Kappa    Std. err.         Z      Prob>Z
------------------------------------------------------------------
  86.67%       69.11%     0.5684       0.0788      7.22      0.0000
```

The `w` weights are given by $1 - |i-j|/(k-1)$, where i and j index the rows of columns of the ratings by the two raters and k is the maximum number of possible ratings. The weighting matrix is printed above the table. Here the rows and columns of the 4×4 matrix correspond to the ratings normal, benign, suspicious, and cancerous.

A weight of 1 indicates that an observation should count as perfect agreement. The matrix has 1s down the diagonals—when both radiologists make the same assessment, they are in agreement. A weight of, say, 0.6667 means that they are in two-thirds agreement. In our matrix, they get that score if they are "one apart"—one radiologist assesses cancer and the other is merely suspicious, or one is suspicious and the other says benign, and so on. An entry of 0.3333 means that they are in one-third agreement, or, if you prefer, two-thirds disagreement. That is the score attached when they are "two apart". Finally, they are in complete disagreement when the weight is zero, which happens only when they are three apart—one says cancer and the other says normal. ◁

▷ Example 3: Weighted kappa, prerecorded weight w2

The other prerecorded weight is `w2`, where the weights are given by $1 - \{(i-j)/(k-1)\}^2$:

```
. kap rada radb, wgt(w2)
Ratings weighted by:
   1.0000   0.8889   0.5556   0.0000
   0.8889   1.0000   0.8889   0.5556
   0.5556   0.8889   1.0000   0.8889
   0.0000   0.5556   0.8889   1.0000
             Expected
Agreement    agreement      Kappa    Std. err.         Z      Prob>Z
------------------------------------------------------------------
  94.77%       84.09%     0.6714       0.1079      6.22      0.0000
```

The `w2` weight makes the categories even more alike and is probably inappropriate here. ◁

▷ Example 4: Weighted kappa, user-defined weights

In addition to using prerecorded weights, we can define our own weights with the `kapwgt` command. For instance, we might feel that suspicious and cancerous are reasonably similar, that benign and normal are reasonably similar, but that the suspicious/cancerous group is nothing like the benign/normal group:

```
. kapwgt xm 1 \ .8 1 \ 0 0 1 \ 0 0 .8 1
. kapwgt xm
 1.0000
 0.8000 1.0000
 0.0000 0.0000 1.0000
 0.0000 0.0000 0.8000 1.0000
```

We name the weights `xm`, and after the weight name, we enter the lower triangle of the weighting matrix, using `\` to separate rows. We have four outcomes, so we continued entering numbers until we had defined the fourth row of the weighting matrix. If we type `kapwgt` followed by a name and nothing else, it shows us the weights recorded under that name. Satisfied that we have entered them correctly, we now use the weights to recalculate kappa:

```
. kap rada radb, wgt(xm)
Ratings weighted by:
   1.0000   0.8000   0.0000   0.0000
   0.8000   1.0000   0.0000   0.0000
   0.0000   0.0000   1.0000   0.8000
   0.0000   0.0000   0.8000   1.0000

            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
------------------------------------------------------------------
  80.47%      52.67%     0.5874      0.0865       6.79      0.0000
```

◁

❑ Technical note

In addition to using weights for weighting the differences in categories, you can specify Stata's traditional weights for weighting the data. In the examples above, we have 85 observations in our dataset—one for each patient. If we only knew the table of outcomes—that there were 21 patients rated normal by both radiologists, etc.—it would be easier to enter the table into Stata and work from it. The easiest way to enter the data is with `tabi`; see [R] **tabulate twoway**.

```
. tabi 21 12 0 0 \ 4 17 1 0 \ 3 9 15 2 \ 0 0 0 1, replace
```

row	col 1	2	3	4	Total
1	21	12	0	0	33
2	4	17	1	0	22
3	3	9	15	2	29
4	0	0	0	1	1
Total	28	38	16	3	85

Pearson chi2(9) = 77.8111 Pr = 0.000

`tabi` reported the Pearson χ^2 for this table, but we do not care about it. The important thing is that, with the `replace` option, `tabi` left the table in memory:

```
. list in 1/5
```

	row	col	pop
1.	1	1	21
2.	1	2	12
3.	1	3	0
4.	1	4	0
5.	2	1	4

The variable `row` is radiologist A's assessment, `col` is radiologist B's assessment, and `pop` is the number so assessed by both. Thus,

```
. kap row col [fweight=pop]
```

Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
63.53%	30.82%	0.4728	0.0694	6.81	0.0000

If we are going to keep these data, the names `row` and `col` are not indicative of what the data reflect. We could type (see **[U] 12.6 Dataset, variable, and value labels**)

```
. rename row rada
. rename col radb
. label var rada "Radiologist A's assessment"
. label var radb "Radiologist B's assessment"
. label define assess 1 normal 2 benign 3 suspect 4 cancer
. label values rada assess
. label values radb assess
. label data "Altman, page 403"
```

`kap`'s `tab` option, which can be used with or without weighted data, shows the table of assessments:

```
. kap rada radb [fweight=pop], tab
```

Radiologist A's assessment	Radiologist B's assessment				
	normal	benign	suspect	cancer	Total
normal	21	12	0	0	33
benign	4	17	1	0	22
suspect	3	9	15	2	29
cancer	0	0	0	1	1
Total	28	38	16	3	85

Agreement	Expected agreement	Kappa	Std. err.	Z	Prob>Z
63.53%	30.82%	0.4728	0.0694	6.81	0.0000

❑

❑ Technical note

You have data on individual patients. There are two raters, and the possible ratings are 1, 2, 3, and 4, but neither rater ever used rating 3:

```
. use https://www.stata-press.com/data/r17/rate2no3, clear
. tabulate ratera raterb
```

Rater A	Rater B 1	2	4	Total
1	6	4	3	13
2	5	3	3	11
4	1	1	26	28
Total	12	8	32	52

Here `kap` would determine that the ratings are from the set $\{1, 2, 4\}$ because those were the only values observed. `kap` would expect a user-defined weighting matrix to be 3×3, and if it were not, `kap` would issue an error message. In the formula-based weights, the calculation would be based on $i, j = 1, 2, 3$ corresponding to the three observed ratings $\{1, 2, 4\}$.

Specifying the `absolute` option would clarify that the ratings are 1, 2, 3, and 4; it just so happens that rating 3 was never assigned. If a user-defined weighting matrix were also specified, `kap` would expect it to be 4×4 or larger (larger because we can think of the ratings being 1, 2, 3, 4, 5, ... and it just so happens that ratings 5, 6, ... were never observed, just as rating 3 was not observed). In the formula-based weights, the calculation would be based on $i, j = 1, 2, 4$.

```
. kap ratera raterb, wgt(w)
Ratings weighted by:
   1.0000   0.5000   0.0000
   0.5000   1.0000   0.5000
   0.0000   0.5000   1.0000
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
------------------------------------------------------------------
  79.81%      57.17%     0.5285      0.1169       4.52       0.0000
. kap ratera raterb, wgt(w) absolute
Ratings weighted by:
   1.0000   0.6667   0.0000
   0.6667   1.0000   0.3333
   0.0000   0.3333   1.0000
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
------------------------------------------------------------------
  81.41%      55.08%     0.5862      0.1209       4.85       0.0000
```

If all conceivable ratings are observed in the data, specifying `absolute` makes no difference. For instance, if rater A assigns ratings $\{1, 2, 4\}$ and rater B assigns $\{1, 2, 3, 4\}$, the complete set of assigned ratings is $\{1, 2, 3, 4\}$, the same that `absolute` would specify. Without `absolute`, it makes no difference whether the ratings are coded $\{1, 2, 3, 4\}$, $\{0, 1, 2, 3\}$, $\{1, 7, 9, 100\}$, $\{0, 1, 1.5, 2.0\}$, or otherwise.

❑

More than two raters

For more than two raters, the mathematics are such that the two raters are not considered unique. For instance, if there are three raters, there is no assumption that the three raters who rate the first subject are the same as the three raters who rate the second. Although we call this the "more than two raters" case, it can be used with two raters when the raters' identities vary.

The nonunique rater case can be usefully broken down into three subcases: 1) there are two possible ratings, which we will call positive and negative; 2) there are more than two possible ratings, but the number of raters per subject is the same for all subjects; and 3) there are more than two possible ratings, and the number of raters per subject varies. `kappa` handles all these cases. To emphasize that there is no assumption of constant identity of raters across subjects, the variables specified contain counts of the number of raters rating the subject into a particular category.

Jacob Cohen (1923–1998) was born in New York City. After studying psychology at City College of New York and New York University, he worked as a medical psychologist until 1959 when he became a full professor in the Department of Psychology at New York University. He made many contributions to research methods, including the kappa measure. He persistently emphasized the value of multiple regression and the importance of power and of measuring effects rather than testing significance.

▷ Example 5: Two ratings

Fleiss, Levin, and Paik (2003, 612) offer the following hypothetical ratings by different sets of raters on 25 subjects:

Subject	No. of raters	No. of pos. ratings	Subject	No. of raters	No. of pos. ratings
1	2	2	14	4	3
2	2	0	15	2	0
3	3	2	16	2	2
4	4	3	17	3	1
5	3	3	18	2	1
6	4	1	19	4	1
7	3	0	20	5	4
8	5	0	21	3	2
9	2	0	22	4	0
10	4	4	23	3	0
11	5	5	24	3	3
12	3	3	25	2	2
13	4	4			

We have entered these data into Stata, and the variables are called `subject`, `raters`, and `pos`. `kappa`, however, requires that we specify variables containing the number of positive ratings and negative ratings, that is, `pos` and `raters-pos`:

```
. use https://www.stata-press.com/data/r17/p612
. generate neg = raters-pos
. kappa pos neg
Two-outcomes, multiple raters:
        Kappa        Z        Prob>Z
-------------------------------------
       0.5415     5.28       0.0000
```

We would have obtained the same results if we had typed `kappa neg pos`.

◁

▷ Example 6: More than two ratings, constant number of raters, kappa

Each of 10 subjects is rated into one of three categories by five raters (Fleiss, Levin, and Paik 2003, 615):

```
. use https://www.stata-press.com/data/r17/p615, clear
. list
```

	subject	cat1	cat2	cat3
1.	1	1	4	0
2.	2	2	0	3
3.	3	0	0	5
4.	4	4	0	1
5.	5	3	0	2
6.	6	1	4	0
7.	7	5	0	0
8.	8	0	4	1
9.	9	1	0	4
10.	10	3	0	2

We obtain the kappa statistic:

```
. kappa cat1-cat3
```

Outcome	Kappa	Z	Prob>Z
Category 1	0.2917	2.92	0.0018
Category 2	0.6711	6.71	0.0000
Category 3	0.3490	3.49	0.0002
combined	0.4179	5.83	0.0000

The first part of the output shows the results of calculating kappa for each of the categories separately against an amalgam of the remaining categories. For instance, the `cat1` line is the two-rating kappa, where positive is `cat1` and negative is `cat2` or `cat3`. The test statistic, however, is calculated differently (see *Methods and formulas*). The combined kappa is the appropriately weighted average of the individual kappas. There is considerably less agreement about the rating of subjects into the first category than there is for the second.

◁

▷ Example 7: More than two ratings, constant number of raters, kap

Now, suppose that we have the same data as in the previous example but that the data are organized differently:

```
. use https://www.stata-press.com/data/r17/p615b
. list
```

	subject	rater1	rater2	rater3	rater4	rater5
1.	1	1	2	2	2	2
2.	2	1	1	3	3	3
3.	3	3	3	3	3	3
4.	4	1	1	1	1	3
5.	5	1	1	1	3	3
6.	6	1	2	2	2	2
7.	7	1	1	1	1	1
8.	8	2	2	2	2	3
9.	9	1	3	3	3	3
10.	10	1	1	1	3	3

Here we would use `kap` rather than `kappa` because the variables record ratings for each rater.

```
. kap rater1 rater2 rater3 rater4 rater5
There are 5 raters per subject:
```

Outcome	Kappa	Z	Prob>Z
1	0.2917	2.92	0.0018
2	0.6711	6.71	0.0000
3	0.3490	3.49	0.0002
combined	0.4179	5.83	0.0000

It does not matter which rater is which when there are more than two raters.

◁

▷ Example 8: More than two ratings, varying number of raters, kappa

In this unfortunate case, kappa can be calculated, but there is no test statistic for testing against $\kappa > 0$. We do nothing differently—`kappa` calculates the total number of raters for each subject, and, if it is not a constant, `kappa` suppresses the calculation of test statistics.

```
. use https://www.stata-press.com/data/r17/rvary
. list
```

	subject	cat1	cat2	cat3
1.	1	1	3	0
2.	2	2	0	3
3.	3	0	0	5
4.	4	4	0	1
5.	5	3	0	2
6.	6	1	4	0
7.	7	5	0	0
8.	8	0	4	1
9.	9	1	0	2
10.	10	3	0	2

```
. kappa cat1-cat3
```

Outcome	Kappa	Z	Prob>Z
Category 1	0.2685	.	.
Category 2	0.6457	.	.
Category 3	0.2938	.	.
combined	0.3816	.	.

```
Note: Number of ratings per subject vary; cannot calculate test
      statistics.
```

▷ Example 9: More than two ratings, varying number of raters, kap

This case is similar to the previous example, but the data are organized differently:

```
. use https://www.stata-press.com/data/r17/rvary2
. list
```

	subject	rater1	rater2	rater3	rater4	rater5
1.	1	1	2	2	.	2
2.	2	1	1	3	3	3
3.	3	3	3	3	3	3
4.	4	1	1	1	1	3
5.	5	1	1	1	3	3
6.	6	1	2	2	2	2
7.	7	1	1	1	1	1
8.	8	2	2	2	2	3
9.	9	1	3	.	.	3
10.	10	1	1	1	3	3

Here we specify kap instead of kappa because the variables record ratings for each rater.

```
. kap rater1-rater5
There are between 3 and 5 (median = 5.00) raters per subject:
```

Outcome	Kappa	Z	Prob>Z
1	0.2685	.	.
2	0.6457	.	.
3	0.2938	.	.
combined	0.3816	.	.

```
Note: Number of ratings per subject vary; cannot calculate test
      statistics.
```

Stored results

kap and kappa store the following in r():

Scalars

r(N)	number of subjects (kap only)	r(kappa)	kappa
r(prop_o)	observed proportion of agreement (kap only)	r(z)	z statistic
r(prop_e)	expected proportion of agreement (kap only)	r(se)	standard error for kappa statistic

Methods and formulas

The kappa statistic was first proposed by Cohen (1960). The generalization for weights reflecting the relative seriousness of each possible disagreement is due to Cohen (1968). The analysis-of-variance approach for $k = 2$ and $m \geq 2$ is due to Landis and Koch (1977b). See Altman (1991, 403–409) or Dunn (2000, chap. 2) for an introductory treatment and Fleiss, Levin, and Paik (2003, chap. 18) for a more detailed treatment. All formulas below are as presented in Fleiss, Levin, and Paik (2003). Let m be the number of raters, and let k be the number of rating outcomes.

Methods and formulas are presented under the following headings:

kap: $m = 2$
kappa: $m > 2$, $k = 2$
kappa: $m > 2$, $k > 2$

kap: m = 2

Define w_{ij} ($i = 1, \ldots, k$ and $j = 1, \ldots, k$) as the weights for agreement and disagreement (wgt()), or, if the data are not weighted, define $w_{ii} = 1$ and $w_{ij} = 0$ for $i \neq j$. If wgt(w) is specified, $w_{ij} = 1 - |i - j|/(k - 1)$. If wgt(w2) is specified, $w_{ij} = 1 - \{(i - j)/(k - 1)\}^2$.

The observed proportion of agreement is

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}$$

where p_{ij} is the fraction of ratings i by the first rater and j by the second. The expected proportion of agreement is

$$p_e = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i\cdot} p_{\cdot j}$$

where $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$.

Kappa is given by $\widehat{\kappa} = (p_o - p_e)/(1 - p_e)$.

The standard error of $\widehat{\kappa}$ for testing against 0 is

$$\widehat{s}_0 = \frac{1}{(1-p_e)\sqrt{n}} \left(\left[\sum_i \sum_j p_{i\cdot} p_{\cdot j} \{ w_{ij} - (\overline{w}_{i\cdot} + \overline{w}_{\cdot j}) \}^2 \right] - p_e^2 \right)^{1/2}$$

where n is the number of subjects being rated, $\overline{w}_{i\cdot} = \sum_j p_{\cdot j} w_{ij}$, and $\overline{w}_{\cdot j} = \sum_i p_{i\cdot} w_{ij}$. The test statistic $Z = \widehat{\kappa}/\widehat{s}_0$ is assumed to be distributed $N(0,1)$.

kappa: m > 2, k = 2

Each subject i, $i = 1, \ldots, n$, is found by x_i of m_i raters to be positive (the choice as to what is labeled positive is arbitrary).

The overall proportion of positive ratings is $\overline{p} = \sum_i x_i/(n\overline{m})$, where $\overline{m} = \sum_i m_i/n$. The between-subjects mean square is (approximately)

$$B = \frac{1}{n} \sum_i \frac{(x_i - m_i \overline{p})^2}{m_i}$$

and the within-subject mean square is

$$W = \frac{1}{n(\overline{m} - 1)} \sum_i \frac{x_i (m_i - x_i)}{m_i}$$

Kappa is then defined as

$$\widehat{\kappa} = \frac{B - W}{B + (\overline{m} - 1)W}$$

The standard error for testing against 0 (Fleiss and Cuzick 1979) is approximately equal to and is calculated as

$$\widehat{s}_0 = \frac{1}{(\overline{m} - 1)\sqrt{n\overline{m}_H}} \left\{ 2(\overline{m}_H - 1) + \frac{(\overline{m} - \overline{m}_H)(1 - 4\overline{p}\overline{q})}{\overline{m}\overline{p}\overline{q}} \right\}^{1/2}$$

where $\overline{m}_H$ is the harmonic mean of m_i and $\overline{q} = 1 - \overline{p}$.

The test statistic $Z = \widehat{\kappa}/\widehat{s}_0$ is assumed to be distributed $N(0,1)$.

kappa: m > 2, k > 2

Let x_{ij} be the number of ratings on subject i, $i = 1, \ldots, n$, into category j, $j = 1, \ldots, k$. Define $\overline{p}_j$ as the overall proportion of ratings in category j, $\overline{q}_j = 1 - \overline{p}_j$, and let $\widehat{\kappa}_j$ be the kappa statistic given above for $k = 2$ when category j is compared with the amalgam of all other categories. Kappa is

$$\overline{\kappa} = \frac{\sum_j \overline{p}_j \overline{q}_j \widehat{\kappa}_j}{\sum_j \overline{p}_j \overline{q}_j}$$

(Landis and Koch 1977b). In the case where the number of raters per subject, $\sum_j x_{ij}$, is a constant m for all i, Fleiss, Nee, and Landis (1979) derived the following formulas for the approximate standard errors. The standard error for testing $\widehat{\kappa}_j$ against 0 is

$$\widehat{s}_j = \left\{ \frac{2}{nm(m-1)} \right\}^{1/2}$$

and the standard error for testing $\overline{\kappa}$ is

$$\overline{s} = \frac{\sqrt{2}}{\sum_j \overline{p}_j \overline{q}_j \sqrt{nm(m-1)}} \left\{ \left(\sum_j \overline{p}_j \overline{q}_j \right)^2 - \sum_j \overline{p}_j \overline{q}_j (\overline{q}_j - \overline{p}_j) \right\}^{1/2}$$

References

Altman, D. G. 1991. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC.

Boyd, N. F., C. Wolfson, M. Moskowitz, T. Carlile, M. Petitclerc, H. A. Ferri, E. Fishell, A. Gregoire, M. Kiernan, J. D. Longley, I. S. Simor, and A. B. Miller. 1982. Observer variation in the interpretation of xeromammograms. *Journal of the National Cancer Institute* 68: 357–363. https://doi.org/10.1093/jnci/68.3.357.

Campbell, M. J., D. Machin, and S. J. Walters. 2007. *Medical Statistics: A Textbook for the Health Sciences*. 4th ed. Chichester, UK: Wiley.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.

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

Cox, N. J. 2006. Assessing agreement of measurements and predictions in geomorphology. *Geomorphology* 76: 332–346. https://doi.org/10.1016/j.geomorph.2005.12.001.

Dunn, G. 2000. *Statistics in Psychiatry*. London: Arnold.

Fleiss, J. L., and J. Cuzick. 1979. The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement* 3: 537–542. https://doi.org/10.1177/014662167900300410.

Fleiss, J. L., B. Levin, and M. C. Paik. 2003. *Statistical Methods for Rates and Proportions*. 3rd ed. New York: Wiley.

Fleiss, J. L., J. C. M. Nee, and J. R. Landis. 1979. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin* 86: 974–977. http://doi.org/10.1037/0033-2909.86.5.974.

Klein, D. 2018. Implementing a general framework for assessing interrater agreement in Stata. *Stata Journal* 18: 871–901.

Landis, J. R., and G. G. Koch. 1977a. The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174. http://doi.org/10.2307/2529310.

———. 1977b. A one-way components of variance model for categorical data. *Biometrics* 33: 671–679. http://doi.org/10.2307/2529465.

Reichenheim, M. E. 2004. Confidence intervals for the kappa statistic. *Stata Journal* 4: 421–428.

Shrout, P. E. 2001. Jacob Cohen (1923–1998). *American Psychologist* 56: 166. https://doi.org/10.1037/0003-066X.56.2.166.